



Aljameel, Sumayh Saud (2018) Development of an Arabic conversational intelligent tutoring system for education of children with autism spectrum disorder. Doctoral thesis (PhD), Manchester Metropolitan University.

Downloaded from: <https://e-space.mmu.ac.uk/622633/>

Usage rights: Creative Commons: Attribution-Noncommercial-No Derivative Works 4.0

Please cite the published version

<https://e-space.mmu.ac.uk>

Development of an Arabic Conversational Intelligent Tutoring System for Education of Children with Autism Spectrum Disorder

Sumayh Saud Aljameel

A thesis submitted in partial fulfilment of the requirements of
the Manchester Metropolitan University for the degree of
Doctor of Philosophy

School of Computing, Mathematics and Digital Technology
Manchester Metropolitan University

December 2018

Abstract

Children with Autism Spectrum Disorder (ASD) are affected in different degrees in terms of their level of intellectual ability. Some people with Asperger syndrome or high functioning autism are very intelligent academically but they still have difficulties in social and communication skills. In recent years, many of these pupils are taught within mainstream schools. However, the process of facilitating their learning and participation remains a complex and poorly understood area of education. Although many teachers in mainstream schools are firmly committed to the principles of inclusive education, they do not feel that they have the necessary training and support to provide adequately for pupils with ASD. One solution for this problem is to use a virtual tutor to supplement the education of pupils with ASD in mainstream schools. This thesis describes research to develop a Novel Arabic Conversational Intelligent Tutoring System (CITS), called LANA, for children with ASD, which delivers topics related to the science subject by engaging with the user in Arabic language. The Visual, Auditory, and Kinaesthetic (VAK) learning style model is used in LANA to adapt to the children's learning style by personalising the tutoring session. Development of an Arabic Conversational Agent has many challenges. Part of the challenge in building such a system is the requirement to deal with the grammatical features and the morphological nature of the Arabic language. The proposed novel architecture for LANA uses both pattern matching (PM) and a new Arabic short text similarity (STS) measure to extract facts from user's responses to match rules in scripted conversation in a particular domain (Science). In this research, two prototypes of an Arabic CITS were developed (LANA-I) and (LANA-II). LANA-I was developed and evaluated with 24 neurotypical children to evaluate the effectiveness and robustness of the system engine. LANA-II was developed to enhance LANA-I by addressing spelling mistakes and words variation with prefix and suffix. Also in LANA-II, TEACCH method was added to the user interface to adapt the tutorial environment to the autistic students learning, and the knowledge base was expanded by adding a new tutorial. An evaluation methodology and experiment were designed to evaluate the enhanced components of LANA-II architecture. The results illustrated a statistically significant impact on the effectiveness of LANA-II engine when compared to LANA-I. In addition, the results indicated a statistically significant improvement on the autistic students learning gain with adapting to their learning styles indicating that LANA-II can be adapted to autistic children's learning styles and enhance their learning.

Table of Contents

Chapter 1 – Introduction	1
1.1 Overview	1
1.2 Research aim	1
1.3 Research Objectives	2
1.4 Research Questions	2
1.5 Background and Motivation	3
1.6 Research contributions	5
1.7 Thesis Structure	6
Chapter 2 – Literature Review	8
2.1 Overview	8
2.2 Autism Spectrum Disorder	8
2.3 Types of Autism	9
2.4 Inclusion of Autistic students in mainstream schools	10
2.5 Interventions	10
2.5.1 TEACCH Method Intervention	11
2.5.2 Technology-Based Interventions	12
2.6 Learning styles	14
2.6.1 VAK learning style	15
2.6.2 Autism and learning style	15
2.6.3 Integrating learning styles into tutoring system	16
2.7 Conversational Agents	16
2.7.1 Goal Orientated Conversational Agents (GO-CA)	18
2.7.2 Conversational Agent Approach	18
2.7.2.1 Natural language processing (NLP)	18
2.7.2.2 Text-based Pattern Matching (PM)	19
2.7.2.3 Sentence similarity	20
2.7.3 Conversational Agents in Arabic Language	26
2.7.4 Challenges within the Arabic Language	27
2.7.5 Conversational Agent Knowledge Base Development	28
2.7.6 Conversational Agent Evaluation	29
2.8 Intelligent Tutoring Systems	30
2.8.1 Architecture of ITS System	31
2.8.2 Conversational Intelligent Tutoring Systems	32
2.8.3 Evaluation of CITS	33
2.9 Chapter Summary	33
Chapter 3 – Research Methodology and Procedure	35

3.1 Overview	35
3.2 Research Design.....	35
3.2.1 Development of LANA CITS	36
3.2.2 Evaluation of LANA CITS	37
3.3 Data collection and analysis methodology of the research	39
3.3.1 Triangulation.....	39
3.4 Ethical consideration.....	40
3.5 Chapter summary	40
Chapter 4 - Methodology for Developing the Arabic CITS (LANA-I)	42
4.1 Overview	42
4.2 LANA-I Architecture Overview	43
4.2.1 Phase 1: Designing an Arabic CA architecture	43
4.2.2 Phase 2: Designing the ITS architecture	45
4.3 Methodology of Implementing LANA CITS.....	45
4.4 Component 1: Implementing the LANA-I Knowledge Base.....	47
4.4.1 The Domain	47
4.4.2 Arabic CITS knowledge engineering.....	47
4.4.3 Arabic scripting language	49
4.4.4 Methodology of Scripting Arabic CA for LANA-I	50
4.4.5 VAK learning style questionnaire	52
4.4.6 Mapping the tutorial into VAK learning style model:	52
4.4.7 Log File.....	54
4.5 Component 2: Designing and Implementing the Arabic CA	55
4.5.1 Controller	55
4.5.2 Conversational Agent Manager	55
4.5.3 Conversational Agent Engine	56
4.5.4 LANA CA Engine Workflow	58
4.6 Component 3: (LANA-I) Intelligent Tutoring System (ITS) Implementation	59
4.6.1 Graphical User Interface	60
4.6.2 ITS Manager	61
4.7 LANA-I Workflow	61
4.8 Chapter summary	65
Chapter 5 – LANA-I Evaluation Methodology and Results.....	66
5.1 Overview.....	66
5.2 (LANA-I) Evaluation Methodology	67
5.2.1 Hypothesis.....	67
5.2.2 Subjective user evaluation	68

5.2.3 Objective user evaluation.....	69
5.3 Design of evaluation metrics.....	69
5.3.1 Experimental Methods for Subjective and Objective Evaluation	70
5.4. Participants and Ethical Consideration	71
5.5. Participants Interaction	72
5.5.1 Experiment 1: LANA-I Tutoring with and without adapting to VAK learning Style	72
5.5.2 Experiment 2 – LANA-I CA System Robustness.....	72
5.6 Experimental Data Analysis.....	73
5.7 Data Normality Testing.....	73
5.7.1 Wilcoxon Signed Ranks Test.....	73
5.7.2 Mann-Whitney U test.....	74
5.7.3 Chi-Square Test of Independence	74
5.8 Experiment 1: results and discussion	75
5.9 Experiment 2: results and discussion	82
5.10 Discussion	86
5.11 Chapter Summary	88
Chapter 6 - LANA-II with Improved Architecture	90
6.1 Overview	90
6.2 Morphologic Pre-processing	91
6.3 Cosine similarity	92
6.4 Knowledge base expansion.....	96
6.4.1 Tutorial 2: Electrical circuit and the magnetism	97
6.5 Mapping TEACCH method to the tutorial.....	98
6.5.1 Physical Structure	99
6.5.2 Work System.....	99
6.5.3 Daily Schedule	99
6.5.4 Visual Cues and Instruction	100
6.6 Updated Architecture Diagram	101
6.7 Scripting tool.....	103
6.8 Chapter Summary	103
Chapter 7 - LANA-II Evaluation Methodology and Results	105
7.1 Overview	105
7.2 Experimental Design.....	106
7.2.1 Ethical considerations	106
7.2.2 Hypothesis.....	107
7.2.3 Designing of evaluation metrics.....	107

7.3 Experimental Methodology	111
7.4 Experimental Data Analysis.....	112
7.5 Experiment 1: Adapting the VAK model to the autistic students learning Style.....	112
7.5.1 Participant	112
7.5.2 Experiment 1 Methodology	113
7.5.3 Experiment 1: Results and Discussion.....	113
7.5.3.1 Pre-test and Post-test – Results and Discussion.....	114
7.5.3.2 Adapting VAK model to autistics learning style (Questionnaire) – Results and Discussion	117
7.5.4 Experiment 1 Conclusion.....	122
7.6 Experiment 2 – LANA-II engine effectiveness	123
7.6.1 Participants.....	123
7.6.2 Experiment 2 Methodology	123
7.6.3 Experiment 2 Results and Discussion.....	124
7.6.3.1 Log file analysis.....	124
7.6.3.2 Objective data analysis from old and new data sets.....	125
7.6.3.3 Number of unrecognised utterances.....	125
7.6.3.4 Number of correct responses.....	126
7.6.3.5 Number of incorrect responses.....	126
7.6.3.6 Number of utterances have corrected response using STS algorithm.....	127
7.6.3.7 Number of rules fired.....	128
7.6.3.8 Number of utterances using Similarity Strength.....	128
7.6.3.9 Time taken to complete the tutorial	129
7.6.3.10 Analysis of questionnaire data	129
7.6.4 Comparative descriptive analysis of data from autistic participants and neurotypical participants	132
7.6.4.1 Duration of conversation whilst engaged with tutorial	132
7.6.4.2 Number of unrecognised utterances.....	133
7.6.4.3 Number of correct responses.....	133
7.6.4.4 Number of incorrect responses.....	134
7.6.4.5 Number of utterances using Similarity Strength.....	135
7.7 Case Study	135
7.7.1 Participants.....	135
7.7.2 Ethical Considerations	136
7.7.3 Data Collection Methods	136
7.7.3.1 Interviewing	137
7.7.3.2 Participant Observation.....	139
7.7.4 Data Analysis	139

7.7.5 Participant's interaction	140
7.7.6 Case Study: Results and discussion	140
7.8 Discussion	143
7.9 Chapter Summary	147
Chapter 8 – Conclusion and Future Work	148
8.1 Overview	148
8.2 Research Contribution	150
8.2.1 Method for developing LANA CA engine.....	150
8.2.2 Method for developing an Arabic scripting language for LANA CITS	151
8.2.3 Method for adapting the tutorial to the VAK learning style	151
8.2.4 Method for designing GUI using TEACCH method	151
8.2.5 Framework for CITS evaluation	152
8.3 Future Work	152
8.3.1 Semantic similarity	152
8.3.2 Knowledge base expansion.....	152
8.3.3 Automatically predict autistic children learning styles	153
8.3.4 Voice Recognition.....	153
References.....	154
Appendices.....	167
Appendix A – Interview Questions with primary school teachers to develop the tutorial (Arabic version)	168
Appendix B – Screenshot of all GUI	169
Appendix C – Learning style questionnaire (Arabic version)	171
Appendix D – Information sheet for parents	172
Appendix E – Consent Form.....	174
Appendix F – Normality histogram of the first evaluation	175
Appendix G – Interview questions with primary school teachers to enhance the tutorial (Arabic version)	176
Appendix H – Normality Histogram of the second evaluation.....	177
Appendix I – Pre-test and Post test questions (Solar system tutorial)	179
Appendix J – Pre-test and Post-test questions (Tutorial 2: Electrical circuit and the magnetism).....	180
Appendix K – CBR Check Certificate (DBS)	181
Appendix L – Manchester Metropolitan Ethic Approval	182
Appendix M – Author Publications	183

List of Figures

Figure 1.1: Research objectives outline and their related chapters	3
Figure 1.2: Background and motivation.....	5
Figure 4.1: The Arabic CA architecture.....	44
Figure 4.2: LANA-I ITS Architecture	45
Figure 4.3: GO-CA Software Development Methodology (Kaleem et al., 2014)	46
Figure 4.4: LANA-I architecture.....	46
Figure 4.5: The Domain Knowledge base layers	48
Figure 4.6: the solar system and earth models that are used with the Kinaesthetic learning style	51
Figure 4.7: Database Schema.....	52
Figure 4.8: LANA-I GUI, Welcome interface and Tutorial interface.	61
Figure 4.9: LANA-I Workflow	63
Figure 4.10: LANA-I Conversation Flowchart.....	64
Figure 5.1: GQM Model for LANA-I.....	70
Figure 5.2: The solar system model and earth model	85
Figure 6.1: The models that are used with the Kinaesthetic learning style.....	98
Figure 6.2: applying the work system approach of TEACCH method.....	99
Figure 6.3: Applying the scheduling approach of TEACCH method.....	100
Figure 6.4: The activity timeline in the pre-test interface..	100
Figure 6.5: Visual instruction approach of the TEACCH method.....	101
Figure 6.6: LANA-II Architecture and the Utterance Processing.....	102
Figure 6.7: The scripting tool.....	103
Figure 7.1: GQM Model for the second prototype evaluation.....	109
Figure 7.2: Question 3 between Control Group and Experimental Group	119
Figure 7.3: Question 4 analysis bar chart.....	120
Figure 7.4: Question 6 between Control Group and Experimental Group	122
Figure 7.5: Bar Chart questionnaire results between LANA-I and LANA-II.....	131

List of Equations

Equation 1: Cosine Similarity Equation.....	58
Equation 2: effect size calculation	74
Equation 3: Learning gain.....	75
Equation 4: Relative Learning Gain.....	79

List of Tables

Table 4.1: scripting language	51
Table 4.2: Mapping the Tutorial questions to VAK model	53
Table 4.3: example of matching the utterance with the patterns script.....	57
Table 4.4: Cosine Similarity Example	58
Table 4.5: an example of calculating the similarity of two utterances, 1 means the two words are identical, 0 means not.	59
Table 5.1: User Feedback Questionnaire (Translated version)	69
Table 5.2: Subjective Evaluation Metrics	71
Table 5.3: Objective Evaluation Metrics	71
Table 5.4: Pre-test and Post-test scores in Control Group	76
Table 5.5: Pre-test and Post-test scores in Experimental Group	76
Table 5.6: The difference between the mean of the pre-test and post-test scores in the Control Group	76
Table 5.7: The P Value between the means from Table 5.6	77
Table 5.8: The difference between the mean of the pre-test and post-test scores in the Experimental Group.....	77
Table 5.9: The P Value between the means from Table 5.8	77
Table 5.10: Mann-Whitney U test for the learning gain between Control Group and Experimental Group.....	78
Table 5.11: Mann-Whitney U test for the relative learning gain between Control Group and Experimental Group.....	79
Table 5.12: Crosstab * VAK - Question 3 (I think I remember what LANA has taught me in the tutorial)	80
Table 5.13: Chi-Square Tests	80
Table 5.14: Testing the effect size	81
Table 5.15: Crosstab * using VAK - Question 4 (I think I can learn better with adapting to my learning style).....	81
Table 5.16: The differences between the mean ranks between the three learning styles (V: visual, A: auditory, K: kinaesthetic)	81
Table 5.17: The P-Values between the means from table 5.15.....	82
Table 5.18: Excerpt from LANA Log file transcript	82
Table 5.19: Data analysis from the log file	85

Table 5.20: Frequency analysis LANA questionnaire	86
Table 6.1: Calculate the similarity between pair of words.....	93
Table 6.2: a pseudo-code representation of the new similarity algorithm.	94
Table 6.3: The number of scripts patterns for the pattern (Four seasons).....	95
Table 6.4: Comparison of the similarity algorithm scores between LANA-I and LANA-II.....	96
Table 6.5: The second Tutorial	97
Table 7.1: The modified feedback Questionnaire (Translated version).....	110
Table 7.2: Subjective Evaluation Metrics	111
Table 7.3: Objective Evaluation Metrics	111
Table 7.4: Experiment 1 metrics	113
Table 7.5: the difference between the mean of the pre-test and post-test scores in the Experimental Group.....	114
Table 7.6: The P Value between the means from Table 7.5	114
Table 7.7: the difference between the mean of the pre and post-test scores in the Control Group	114
Table 7.8: The P Value between the means from Table 7.7	115
Table 7.9: Mann-Whitney U test for the differences between the pre and post scores between Group1 and Group2.	115
Table 7.10: Mann-Whitney U test for the relative learning gain between Control Group and Experimental Group.	116
Table 7.11: the differences between the mean between the three learning style (V: visual, A: auditory, K: kinaesthetic))	117
Table 7.12: The P-Values between the means from table 7.10.....	117
Table 7.13: Crosstab *with and without VAK - Questiona3: I think I remember what LANA-II has taught me in the tutorial.	118
Table 7.14: Chi-Square Test **with and without VAK - Questiona3: I think I remember what LANA has taught me in the tutorial.	118
Table 7.15: The effect size table for question 3..	118
Table 7.16: Crosstab Questiona4: I think I can learn better with adapting to my learning style (Visual – Auditory -Kinaesthetic).	120
Table 7.17: Crosstab *with and without VAK - Question 6: I think I can learn from LANA-II better than learning from my teacher.	121

Table 7.18: Chi-Square Test *with and without VAK – Question 6: I think I can learn from LANA-II better than learning from my teacher.	121
Table 7.19: The effect size table for Question 6.	121
Table 7.20: Experiment 2 metrics	124
Table 7.21: Log file analysis from the second and first prototype.....	125
Table 7.22: Mann-Whitney U test for unrecognised utterances	125
Table 7.23: Mann-Whitney U test for correct responses	126
Table 7.24: Mann-Whitney U test for incorreced responses	126
Table 7.25: Mann-Whitney U test for corrected responses using STS algorithm ...	127
Table 7.26: Mann-Whitney U test for the number of the rules fired	128
Table 7.27: Mann-Whitney U test for the numbers of utterances using similarity strength.	128
Table 7.28: Mann-Whitney U test for the average of the conversation duration.....	129
Table 7.29: Frequency analysis LANA-I questionnaire data results	130
Table 7.30: Frequency analysis LANA-II questionnaire data results	130
Table 7.31: Mean values from evaluation questionnaire LANA-I	130
Table 7.32: Mean values from evaluation questionnaire LANA-II	130
Table 7.33: Mann Whitney test between LANA-I and LANA-II questionnaire data.	131
Table 7.34: Mann-Whitney U test for duration of conversation between the two groups participants	132
Table 7.35: Mann-Whitney U test for the number of unrecognised utterances between the two groups participants	133
Table 7.36: Mann-Whitney U test for the number of correct utterances between the two groups participants	134
Table 7.37: Mann-Whitney U test for the number of incorreced utterances between the two groups participants	134
Table 7.38: Mann-Whitney U test for the numbers of utterances using similarity strength.	135
Table 7.39: The Interview Sheet for the autistic student (translated version).	138
Table 7.40: The Interview Sheet with the teacher (translated version).....	139
Table 7.41: The pre-test and post-test scores for the three participants.....	143

Acknowledgements

يَا رَبِّ لَكَ الْحَمْدُ كَمَا يَنْبَغِي لِجَلَالِ وَجْهِكَ وَلِعَظِيمِ سُلْطَانِكَ

O Allah, for You is all praise as befits the Glory of Your Countenance and the Greatness of Your Might.

All praises and thanks are for **Allah Almighty**, for guidance and the knowledge you have given me. Thank you for giving me the strength and encouragement especially during all the challenging moments in completing this thesis. Indeed everything is from Allah and I am eternally grateful.

I would like to dedicate this thesis to my beloved parents, my father, **Saud Aljameel** and my mother **Norah Alanbari**, for inspiring me throughout my life. They were behind me all the way, constantly encouraging me to complete my study successfully. I greatly value their prayers, support, and love. I am very fortunate to be their daughter. Grateful thanks also go to my father in law, **Suliman Almunif**, my mother in law, **Awatif Almunif**, my brothers and my sisters for their endless encouragement and support.

I gratefully acknowledge the invaluable guidance, sympathetic attitude, moral support, inspiring comments and strong motivation of my supervisor **Dr. James O'Shea**, whose insight made me understand the important and essential aspects of my research.

To my co-supervisor, **Dr. Keeley Crockett**, for always encouraging my effort in my work and for providing insightful feedback during all the stages of the research described in this thesis. Her vision was fundamental in shaping my research and I am very grateful for having had the opportunity to learn from her.

Special thanks to my second co-supervises **Dr. Annabel Latham**, for providing valuable advice, support and guidance throughout all phases of completion of endure.

My sincere appreciation is extended to my advisor **Dr. Mohammed Kaleem** for offering ideas, comments, suggestions and continual support during the entire process of the study.

My most profound gratitude goes to my sponsor, **The Saudi government** represented by the **Imam Abdulrahman bin Faisal University**, Ministry of Education and the

Saudi cultural Bureau in London, for their continuing help and support whilst undertaking this research. It is my pleasure to extend my gratitude to **the Ministry of Education in Saudi Arabia** and **the Ministry of Education in Bahrain**, which cooperated with me through the data collection phase of my research, with a special thanks to all the expert teachers and the participants for their cooperation in the completion of this research.

I have shared numerous happy and sad moments with my friends in Manchester, which can hardly be expressed and acknowledged in words. I would like to thank my friends, **Dr. Albandri, Dr. Noufa, and Dr. Dina**, for their extreme support and kindness during the past three years.

At last, I do not know how to begin with saying thank you to my soul mate, the person that I will be forever indebted to, my dearest husband, **Naif Almunif**. I want to express my appreciation for everything you did for me. Thank you for your constant support, encouragement, sharing and inspiration that I would complete my PhD successfully. Thank you for being patient with me, during the days I was stressed. I greatly value his contribution and deeply appreciate his belief in me. Special thanks to my lovely children, **Yara, Suliman, and Lana**. They were my constant source of happiness and joy; without their smiles I would never have overcome the challenges which my PhD posed.

List of Publications

Aljameel, S. S., O'Shea, J. D., Crockett, K. A., & Latham, A. (2016). A Review of Current Technology-Based Intervention for School Aged Children with Autism Spectrum Disorder. In *Proceedings of SAI Intelligent Systems Conference* (pp. 868-879). Springer.

Aljameel, S. S., O'Shea, J. D., Crockett, K. A. and Latham, A. (2016) *Survey of string similarity approaches and the challenging faced by the Arabic language*. 11th IEEE International Conference on Computer Engineering & Systems, Egypt.

Aljameel, S. S., O'Shea, J. D., Crockett, K. A., Latham, A. (2017) *Development of An Arabic Conversational Intelligent Tutoring System for Education of Children with ASD*. IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications, France.

Aljameel, S. S., O'Shea, J. D., Crockett, K. A., & Latham, A. (2018). Can LANA CITS Support Learning in Autistic Children? A Case Study Evaluation. In *Proceedings of SAI Intelligent Systems Conference* (pp. 374-386). Springer.

Aljameel, S., O'Shea, J., Crockett, K., Latham, A. and Kaleem, M. (2019) LANA-I: An Arabic Conversational Intelligent Tutoring System for Children with ASD. In *Proceedings of Computing Conference*. Springer.

Aljameel, S., O'Shea, J., Crockett, K., Latham, A. and Kaleem, M. (2018) Enhance learning skills for children with ASD in mainstream schools: A Conversational Intelligent Tutoring System. (Under review)

List of Abbreviations

AI	Artificial Intelligence
AS	Asperger Syndrome
ASD	Autism spectrum disorder
CA	Conversational Agent
CITS	Conversational Intelligent Tutoring System
ECA	Embodied Conversational Agent
GO-CA	Goal Orientated Conversational Agents
GQM	Goal Question Metric
GUI	Graphical User Interface
HFA	High Functioning Autism
ITS	Intelligent Tutoring System
KB	Knowledge Base
LCA	Linguistic Conversational Agent
NLP	Natural Language Processing
PM	Pattern Matching
STS	Short Text Similarity
TEACCH	Treatment and Education of Autistic and related Communications Handicapped Children
VAK	Visual, Auditory and Kinaesthetic learning styles model

Chapter 1 – Introduction

1.1 Overview

This thesis presents a Novel Adaptive Arabic Conversational Intelligent Tutoring System (CITS) for children with autism spectrum disorder (ASD) using the Visual, Auditory and Kinaesthetic learning styles model (VAK). This system, known as LANA, implements an Arabic CITS suitable for children with autism aged 10-12 using the (VAK) model in order to answer the research questions stated in section 1.3. The Arabic CITS architecture uses both a pattern matching (PM) algorithm and an Arabic short text similarity (STS) measure to extract facts from user's responses to match rules in scripted conversation in a particular domain (Science). Two tutorials covering the topics (the solar system and electricity) were designed and implemented in accordance with the TEACCH (the Treatment and Education of Autistic and related Communication handicapped Children) model and mapped to the VAK learning style. The mapped tutorials were then incorporated into LANA. This chapter provides the context and motivations behind this research along with a short summary of the research contributions and a brief outline of the thesis structure.

1.2 Research aim

The primary aim of this research is to design and implement an Arabic CITS for Autism called LANA using the VAK model and TEACCH method to enhance autistic students' learning in mainstream schools. The LANA Conversational Agent (CA) architecture used pattern matching (PM) and Arabic short text similarity (STS) to extract facts from user's responses to match rules in scripted conversation in a particular domain (Science). The outcome of the project is the development of an Arabic CITS which is aimed at autistic pupils (10 to 12 years old), who are studying in Arabic mainstream primary schools. Neurotypical children can also use this system as it is generalised for all pupils learning styles. Facilitating the learning of these pupils using traditional teaching within mainstream schools is complex and poorly understood (Waddington and Reed, 2017). This thesis answers research questions through the development and evaluation of two LANA CITS prototypes using the

VAK learning style model to adapt to autistic pupils learning style and improve their learning in mainstream schools.

1.3 Research Objectives

1. Research current education technology for children aged 10-12 who exhibit ASD.
2. Undertake a review of CAs with Arabic Natural Language Processing resources for Arabic short text similarity and assess viability for incorporation in an Arabic CITS.
3. For a chosen domain (Science), design an educational tutoring scenario in accordance with the TEACCH method and use autistic learning styles and adaptation of the tutorial based upon the VAK learning style model.
4. Design and implement architecture for Arabic CITS for appropriate education scenario.
5. Design an experimental methodology to validate the educational tutoring scenario in the Arabic CITS and conduct a study in Arabic speaking countries to evaluate the ability of CITS to adapt to autistic children learning style.

Figure 1 outlines the research objectives, and where in this thesis it is addressed and situated.

1.4 Research Questions

Two general research questions are addressed by this work:

1. Can an Arabic Conversational Intelligent Tutoring System (CITS) be adapted to the learning styles of autistic children?
2. Can an Arabic Conversational Intelligent Tutoring System (CITS) for children with autism enhance their learning?

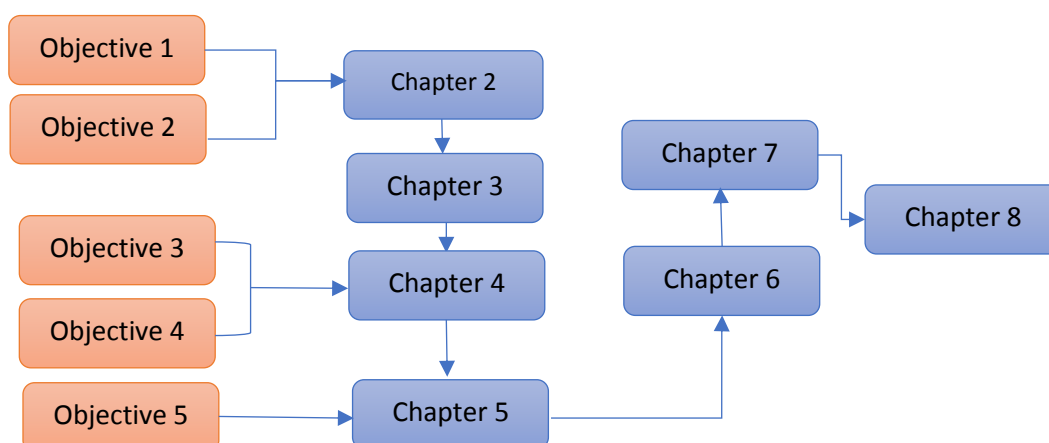


Figure 1.1: Research objectives outline and their related chapters

1.5 Background and Motivation

Autism spectrum disorder (ASD) is a lifelong developmental disability, which affects communication skills, social skills and (repetitive) behaviour. In recent years, many pupils with high-functioning autism or Asperger's have been taught within mainstream schools. However, the process of facilitating their learning and participation remains a complex and poorly understood area of education (Waddington and Reed, 2017). The limited research base in this area indicates that school is a stressful and anxiety-provoking place for many such students with social isolation, loneliness and bullying commonplace (Roberts and Simpson, 2016). Furthermore, although many teachers in mainstream schools are firmly committed to the principles of inclusive education, they do not feel that they have the necessary training and support to provide adequately for students with ASD (Lindsay et al., 2014). Two recent reviews have identified teaching strategies and approaches for students with ASD in mainstream schools as a key 'gap' in the knowledge base for special educational needs provision (Meng, 2017; Barnhart, 2017). One solution is to use a technology-based learning as a virtual tutor to supplement the education of children with ASD in mainstream schools. Several studies reported the benefits of technology-based learning for people with ASD (Grynszpan et al., 2014; Fletcher-Watson, 2014). The virtual tutor offers a structured environment that is preferred by people with autism as they usually like routine and repetitive behaviours (Boucenna et al., 2014). Virtual tutors benefit ASD children who have difficulties with social interactions (Parsons and Cobb, 2011).

Intelligent Tutoring Systems (ITS) include any computer program used in learning that contains intelligent technologies to personalise learning according to individual student's characteristics, such as knowledge of the subject, emotion, and learning style (Dermeval et al., 2017). ITS use Artificial Intelligence methods to provide instructions and feedback to the learners (Perikos et al., 2017). An extension of an ITS is Conversational Intelligent Tutoring Systems (CITS), which use natural language interfaces rather than menus to engage with a learner (Latham et al., 2014). CITSs allow users to explore and learn topics through conversation and to obtain knowledge in a similar manner as in a classroom i.e. through student teacher interaction.

Many pupils with ASD tend to be affected differently as a result of the same symptoms, so it is important to know how each child is impacted on in terms of their cognitive and interaction abilities. Hence, the specific requirements of each child, which is known as learning style, should be identified (D'Elia et al., 2014). Pupils with ASD are more likely to rely on one style of learning, which could be determined by observing the person (Edelson, 2000). VAK theory is considered to be one of the classical learning theories in the educational field (Mackay, 2010). It classifies learners by sensory preferences: Visually (V), Auditory (A), and Kinaesthetic (K) (Miller, 2001). There are evidences for learning styles among ASD pupils but not for a single conclusive learning style for ASD pupils. Some researchers (Banire et al., 2015; O'Connell et al., 2013; Snyder et al., 2012) found that many people with autism are visual learners as they can learn well through watching videos, pictures and movies. Another study (Lisle, 2007) found that autistic people have learning preferences for kinaesthetic, visual, then auditory. Roberts (Roberts, 2010) found that people with autism who are kinaesthetic learners, learn better when "hands-on" activities are included with the teachers taught material to achieve learning. Nevertheless, research that has been conducted with autistic children regarding the impact of learning style is limited and needs further research. The existing research and software applications are not conducive to meeting the individualized needs of pupils with Autism. In addition, there is no Arabic CITS, which adapts to autistic learning styles in order to enhance their learning in a specific domain. In this research, an Arabic CITS was developed to support autistic learning through personalised learning style using the VAK model. The Arabic CITS engages children with science tutorials where the curriculum material is mapped to the VAK model and the

evaluation took place in the UK and Saudi Arabia. The project used a number of evaluation tools such as the learning gain measurement (Kelly and Tangney, 2006; Graesser et al., 2005; Lee et al., 2006), log file, and questionnaire to evaluate the ability of LANA to adapt to the autistic learning styles.

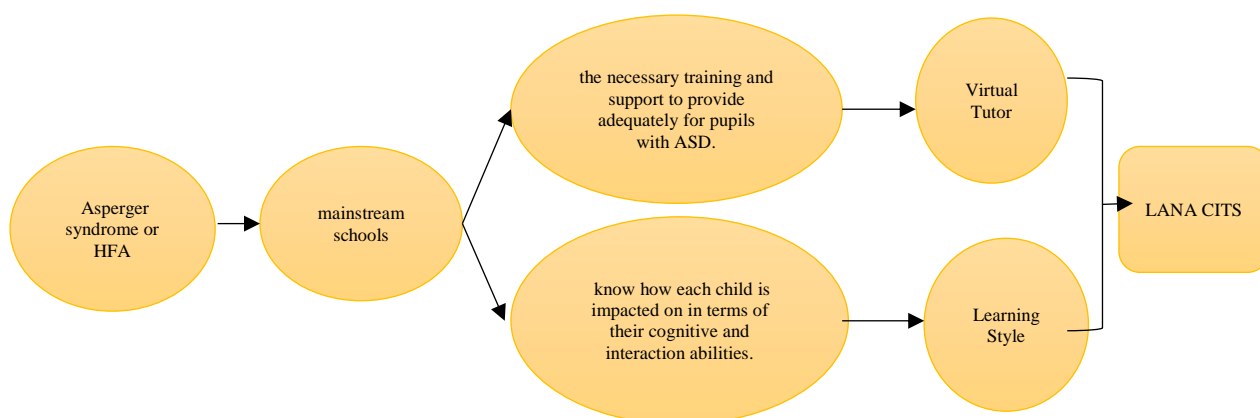


Figure 1.2: Background and motivation.

1.6 Research contributions

The novel contributions of this research are:

1. A Novel architecture for creating an Arabic CITS and a generic development methodology (Chapter 4, Chapter 6).
2. Methodology of designing an educational tutoring scenario, which implements in accordance with the TEACCH method and mapped to the VAK learning style (Chapter 4, Chapter 6).
3. A Novel LANA CA algorithm that reduces the scripting effort by processing the spelling mistakes and addressing the free word order, which is one of the challenges of the Arabic language (Chapter 6).
4. A new evaluation framework to validate an Arabic CITS from the objective and subjective perspective (Chapter 5, Chapter 7).
5. Methodology of developing the first prototype LANA-I and validate the architecture with neurotypical children to evaluate the effectiveness and robustness of the system engine (Chapter 4, Chapter 5).
6. Methodology of developing the enhanced prototype LANA-II and validate the architecture and the adaptive CITS to autistic students' learning styles. (Chapter 6, Chapter 7).

7. Evidence that the Arabic CITS using VAK learning style model can be adapted to autistic children learning styles (Chapter 7).
8. Evidence that the Arabic CITS using VAK learning style model can enhance autistic children learning (Chapter 7).

1.7 Thesis Structure

The research conducted in this thesis is presented over nine chapters. Chapter two details the background review of the existing literature and the current state of research related to the following:

- Technology based intervention for children with ASD.
- The concept of a conversational agent, the limitations of current Arabic CA architecture components and the challenges of the Arabic language.
- The concept of a conversational intelligent tutoring system (CITS), the current state of existing CITS and the limitation of current Arabic CITS.

Chapter three details the research methodology and the theoretical basis of the research derived from the literature chapter. It details the main phases of developing and evaluating LANA CITS and the ethical considerations for this research. In addition the data collection and analysis methodology for LANA experiments is outlined and presented in this chapter.

Chapter four presents the development process adopted to implement an Arabic CITS (called LANA-I). The development process was divided into two phases: the first phase was designing and implementing an Arabic CA, and the second phase was integrating the Arabic Intelligent Tutoring System (ITS) interface with the CA.

Chapter five of the thesis presents the evaluation methodology and results of the empirical experiments carried out to test the LANA-I engine from both objective and subjective aspects. The study was conducted with neurotypical participants from the target age group (10-12) years old whose first language is Arabic. None of the children have been diagnosed as Autistic. It is important to test the quality of the tutorial and the materials, which are used in LANA-I, with the general population before testing it with Autistic. The system robustness and its ability to complete the task must be tested with general population before autistic children in order to avoid any confusion that could distress the autistic child during the tutorial. The results and discussion are

detailed in this chapter to highlight the aspects of the architecture that needed to be strengthened in order to increase the effectiveness of LANA-I, such as the issue of the spelling mistakes made by the users. The results of the pilot study were used to adjust the full-scale experiment to ensure its effectiveness.

Following the evaluation experiments, chapter six presents the further research and development undertaken to overcome the weaknesses of the LANA-I architecture highlighted through the evaluation. This chapter illustrates the amendments made to the existing architecture components as well as the new components added to the architecture of LANA-II such as TEACCH method and an additional tutorial.

Chapter seven presents the evaluation methodology and the results of the experiments conducted to test the second prototype LANA-II. Three experiments were conducted to evaluate LANA-II. The first experiment was conducted with autistic children on the target age group (10-12) years old whose first language is Arabic and have a high functioning autism (HFA) i.e. have no problem with language and no intellectual disabilities. This experiment was conducted to measure the capabilities from different perspectives such as tutoring success, VAK learning style model and user evaluation. The second experiment was conducted with participants on the target age group (10-12) years old whose first language is Arabic. None of the children had been diagnosed as Autistic. This experiment was conducted to test the enhancement made in LANA-II engine by comparing the metrics from LANA-I and the metrics from LANA-II in order to bring to light whether the enhancement made in LANA-II components improve LANA-II's conversation. The third experiment was a case study (qualitative study) with three autistic students and their teachers to test if LANA-II is appropriate as a virtual tutor for Autistic students. The results derived from the three experiments contributed towards answering the main research question.

Chapter eight presents the conclusions drawn from the research findings and discussion. It also outlines the main contributions of the research and provides recommendations for future research.

Chapter 2 – Literature Review

2.1 Overview

Developing a novel of CITS for ASD children is based on principles adopted from research in different key areas, specifically: Autism, Technology-Based Interventions for autistic children, Arabic NLP, Conversational agent architecture with Arabic language, and CITS architectures with Arabic language, and learning style. This chapter provides a background review of existing literature and the current state of research in the related areas:

- Definitions of autism and interventions for children with ASD with a focus on this research project.
- The concepts of conversational agents and the limitations of current Arabic CA architecture components.
- The concepts of conversational intelligent tutoring system (CITS) and the current state of exist CITSs focusing on limitations of current Arabic CITS.
- Learning styles and the impact of using VAK learning style with neurotypical children and autistic children.

2.2 Autism Spectrum Disorder

Autism spectrum disorder (ASD) is classified as a lifelong developmental disability, which affects the communication skills, social skills and repetitive behaviour. Because autism is a spectrum condition, people with autism share specific difficulties, but they will be affected by their condition in different ways. Some autistic people have the ability to live relatively independently whereas others need specialist support to live their lives. In addition, some have a low IQ scores, some have average IQs, and some have over the average. People with autism may also have a sensitivity to sounds, lights, colours, touch, smells and tastes. Autism affects three areas: communication (verbal and non-verbal language), social interaction (understanding and recognition emotions and expression), and repetitive behaviours (adaptation to the environment). Over the past two decades, the cases of autism have increased worldwide. It is reported that more than one out of every 100 people has autism in England, Scotland, Wales and Northern Ireland (TheNationalAutisticSociety, 2015).

Some studies suggest that ASD has genetic causes (O’Roak and State, 2008). However, observations of behavioural aspects are the only way to diagnosis ASD. The average age of diagnosis ASD is over three years old, although other studies recommend that all children should be screened for ASD earlier at 18 to 24 months (Mandell et al., 2005).

Currently there are two separate diagnostic guidelines and classification systems of autism, the American Psychiatric Association Diagnostic and Statistical Manual of Mental Disorders fifth edition (DSM-5), published in May 2013, and the World Health Organization International Classification of Diseases (ICD), which is in its tenth revision (AmericanPsychiatricAssociation, 2013) (Organization, 2004).

There are many features contributing to diagnosis ASD, which usually appear in pre-school children:

- Usually children with autism have a delay in their speech. Some of them have not managed to speak at least 10 words by the age of three, where others do not speak at all. Their speech has a flat monotonous sound. In addition, they tend to repeat sets of words and phrases (Kujala et al., 2013).
- Autistic children may be unable to respond to their name being called, even though they have normal hearing. They are not interested in interacting with other people even with children of similar age. They prefer to play alone and they avoid eye contact. In addition, they communicate with other people using gestures such as pointing or facial expressions (Lord et al., 2013).
- Children with autism have repetitive movements such as rocking back and forth, flapping their hands and flicking their fingers. They tend to play with toys in a repetitive way, prefer to have one routine and are very upset when their routine is changed. Food is another issue with autistic children and they usually dislike a certain food depending on the food texture or colour more than the taste. In older children and teenagers, they have a specific interest in a certain subject or activity (Whitaker et al., 2001).

2.3 Types of Autism

According to (researchautism, 2016), ASD has four main types: (1) Autistic Disorder (classic autism) - People with classic autism usually have significant language delays,

communication, social challenges, and behaviors. (2) Asperger Syndrome (AS): which sometimes called high functioning autism (HFA), they usually function at the higher level of the autism spectrum and have social challenges and different behaviors and interests. However, they typically do not have problems with language or have intellectual disabilities. (3) Childhood Disintegrative Disorder: this is a rare form of autism, which called CDD, disintegrative psychosis, dementia infantilis, or Heller's syndrome. People with CDD develop normally until two years old then they go backwards by losing many of the skills that they had before, such as the ability to talk or walk. (4) Pervasive Developmental Disorder (Not Otherwise Specified): it is a type of autism, which known as PDD(NOS) or atypical autism. Some people do not fit neatly into one of the specific kinds of autism, so their situation is described as PDD(NOS). This research conducted in this project is concerned with high-functioning autism or Asperger.

2.4 Inclusion of Autistic students in mainstream schools

Pupils with Autism Spectrum Disorder (ASD) are affected in different degrees in terms with their level of intellectual ability. People with Asperger syndrome or high functioning autism are very intelligent academically but they still have difficulties with social and communication skills (Bauminger-Zviely et al., 2013). In recent years, many of these pupils have been taught within mainstream schools. However, the process of facilitating their learning and participation remains a complex and poorly understood area of education (Barnard et al., 2000; Davis et al., 2004). Many teachers in mainstream schools encourage the principles of inclusive education, but they do not feel that they have the enough training to teach autistic pupils (Gregor and Campbell, 2001; Robertson et al., 2003). (Davis et al., 2004) reported in their review paper that teaching strategies and approaches for students with autism in mainstream schools are the main key 'gap' in the knowledge base for special educational needs provision.

2.5 Interventions

Interventions for children with ASD focus on training skills, providing learning opportunities, and developing friendly environments to increase their social communication and reduce dysfunctional behaviors. Early detection and early intervention are important to improve long term outcomes. Interventions of children with ASD should be included in preschool/schools, home, and the workplace. Many

approaches are used as intervention for children with ASD including: Picture Exchange Communication System (PECS) as communication intervention (Cagliani et al., 2017), applied behaviour analysis (ABA) as behaviour intervention (Makrygianni et al., 2018), and Treatment and Education of Autistic and related Communications Handicapped Children (TEACCH) as education intervention (Mesibov, 2018).

The research conducted in this thesis and hence this literature focuses on the educational interventions, which typically provide a friendly classroom environment to help the children with ASD to be included in the mainstream schools. Two educational interventions are reviewed below: the TEACCH method, and Technology based interventions.

2.5.1 TEACCH Method Intervention

TEACCH (the Treatment and Education of Autistic and related Communication Handicapped Children) is an education programs approach for people with ASD. This approach is developed by Eric Schopler at the University of North Carolina in 1972 (D'Elia et al., 2014). TEACCH is based on structured teaching technique for autistic students. This program is used in UK schools since 1990 then the National Autistic Society starts collaboration with the Division TEACCH (Mesibov, 2018). The aim of structured teaching is to enhance the teaching methods and the environment by considering the cognitive ability, needs and interest of the autistic students. In addition, the structured teaching aims to organize the classroom and to make the teaching methods and style friendly for autism (Mesibov and Shea, 2010). Many advantages could be gained from using the structured teaching as a method of delivering the curriculum, such as, enhancing and facilitating the teaching and learning process. In addition, it can improve access to the curriculum for many pupils with ASD (Mesibov, 2018).

TEACCH has four main principles. First, physical structure, which is arranging the materials, furniture and general surrounding in order to enhance the environment. Second, work system, which is organising the activity by explaining the activity, the required time to finish this activity, evaluating the student's progress, and what will they gain after completing the activity. Third, daily schedule, which provides all activities and the order in which they will be conducted during the day. This helps

autistic students to understand what they have to do clearly and give them the predictable environment they need. Fourth, visual cues and instruction, which can be used to clarify the tasks, it provides the pupil with ASD with information of how to complete the task because visual skills are usually strong in autistic pupils. Visual information can be very helpful in developing understanding for pupils with ASD. (Panerai et al., 2002) conducted a study to compare two education methods for autistic students, the TEACCH method and the integration programme. The experiment was to apply the TEACCH method to the experimental group, while the control group was integrated in regular schools with a support teacher. The results indicated that the experimental group scores have increased more than the control group scores. In other research (Panerai et al., 2009), a study was conducted over a period of 3 years. This study aimed to test the effectiveness of three different educational approaches with autistic children, which are: the TEACCH method in a residential autism centre (R-TEACCH), the TEACCH method implemented at home and at mainstream schools (NS-TEACCH), and inclusive education in mainstream schools (INSP), in which a nonspecific approach was implemented. The results revealed that the children in R-TEACCH and NS-TEACCH performed better than the children in INSP. In this research, the TEACCH method had positive outcomes in terms of inclusive value. The results indicate that TEACCH and “inclusion” strengthen each other if they are used together. However, inclusion children with ASD without using the TEACCH did not seem to be effective. Mainstream schools need an effort to adopt structured teaching and flexibility with autistic students to create appropriate conditions for an optimal development.

2.5.2 Technology-Based Interventions

Several studies have reported that people with ASD have the ability to deal with technology and computers better than neurotypical people (Putnam and Chong, 2008). Computer’s software offers a structured environment that is preferred by people with autism because of their routine and repetitive behaviours (Murray, 1997). Virtual tutors benefit ASD children, who have difficulties with social interactions (Parsons and Cobb, 2011; Saadatzi et al., 2018; Bozgeyikli et al., 2018). Traditional education using human tutors are a challenge for children with autism whereas virtual tutors can meet the individual children’s needs. For example, some autistic children are

uncomfortable looking at faces, the virtual tutor face looks like a cartoon, and when the child becomes familiar with it, the complexity can be increased. In the same way, the session content can be adjusted depending on the child's current level of need and interest. There is some evidence to suggest that autistic children who were taught by virtual tutors gain more information than who were taught by human tutors (Grynszpan et al., 2008). Many benefits could be gained from virtual tutors, for example, they allow children to learn and practice their skills independently with no pressure on the person who works with them. Consequently, the human tutor will focus on the complex aspects and tasks while the software covers the general topics (Massaro, 2004). Another advantage is that the virtual tutor never gets impatient or tired and learners can practice at any time of the day, which suits with autistic people who have abnormal sleep patterns (Limoges et al., 2005). Regarding developing nonverbal skills, it is reported that virtual tutors can be useful in modelling the learner's behaviours, such as gaze and facial expression. Video modelling is one of the successful examples that is used with many people with autism (Marcus and Wilder, 2009). Technology-based interventions are the most studied intervention for children with ASD. Technology has been used to teach a variety of skills, such as how to recognize and predict emotions (Silver and Oakes, 2001), enhance problem solving (Bernard-Opitz et al., 2001), improve vocabulary (Massaro and Bosseler, 2006), and improve reading and communication skills (Wainer and Ingersoll, 2011).

As a part of the research presented in this thesis, a further literature review was conducted which can be found in (Aljameel et al., 2016). This paper reviewed technologies-based intervention for children with ASD. The review contributes to the literature in three important ways. First, it focused on technologies that are used as an intervention for children with ASD aged 10 – 16 years old. Second, this review concentrated on technologies that enhanced intervention, which targets three cognitive domains: (1) languages and literacy, (2) social skills, and (3) emotion recognition. Third, this review was comprehensive in that it provided information across the participant's number, diagnoses criteria, specific target, context, evaluation methods, description of training, and results. Nineteen articles were reviewed in this paper (Appendix M), and the results indicated that there is a sufficient improvement in learning for children with ASD within the computerised training paradigm. However,

many of the existing technologies have limited capabilities in their performance and thus limit the success in the clinical approach of children with ASD:

- Due to difficulties in recruiting participants, the sample size was small in all studies, and most have no control group. Future research should increase the participants number and to involve a control group without Cognitive Behavioural Treatment (CBT) and a control group with CBT and without collaborative technology.
- Participants included in this review were diagnosed with ASD, and have differences in functional abilities. Future research should examine the effect of participant characteristics on program results. For example, a program that teaches children with ASD advanced language may not be appropriate for children with ASD who uses one to two words to communicate.
- Most studies did not assess whether the participants demonstrate the target behaviors in daily life or not. All the participants performed their tasks in the system with improved self-efficacy after the intervention.

2.6 Learning styles

A learning style is “a particular way in which an individual learns” (Butler, 1987). Different students prefer to learn in different ways. Some may learn quickly through texts and reading, others may prefer images or experiments. Different learning styles provide interventions that are tailored to individual needs.

The debate about learning styles is widely known in educational circles. There has been some critique on learning styles. Some critics believe that it is hard to use learning styles in a face-to-face environment such as classrooms. They believe that it is not possible for a teacher to consider each student’s individual learning style and it is costly and time-consuming (Coffield et al., 2004). This issue could be solved by splitting classrooms by learning styles or to use the virtual learning systems. Many researchers approved that virtual learning systems can adapt to individual students’ learning styles (Latham et al., 2014; Truong, 2016).

Using a suitable learning style will enable learners to be more confident in learning and optimize their learning paths (Herod, 2004). For teachers, learning styles will help them to match suitable learning materials to different groups of students at the

appropriate stage of the learning process (Stash, 2007). In a survey research conducted by (Howard-Jones, 2014) with British teachers, the results indicated that 93% of the teachers have agreed that individuals learn better when they receive information in their preferred learning style such as visual, auditory or kinaesthetic.

There are many models of learning styles from the literature such as Kolb (Kolb, 1976), Dunn and Dunn (Dunn and Dunn, 1974), Honey and Mumford (Honey and Mumford, 2006), Felder- Silverman (Felder and Silverman, 1988), and Visual-Auditory-Kinesthetic (VAK) learning style model (Fleming, 2001). In our work, VAK learning style is used because it is the most widely-used, simple, and suitable for children (Yahya and Noor, 2015).

2.6.1 VAK learning style

VAK theory is considered to be one of the classical learning theories in the educational field (Mackay, 2010). It classifies learners by sensory preferences: Visually (V), Auditory (A), and Kinaesthetic (K). (Smith, 1998) reported in his book the importance of using the VAK learning style in primary school to accelerate the students' learning. Smith reported that accelerated learning framework is to engage students to learn and think using three techniques, which are visual, auditory, kinaesthetic (VAK). He suggested applying a modification in school such as mind-mapping, musical stimulation and hands on activities in order to map the teaching style with VAK learning style. (Akplotsyi and Mahdjoubi, 2011) studied the effects of learning styles on engaging children in school projects. Based on a child friendly-customised VAK learning style preferences questionnaire, they classified children's learning preferences into three sensory modalities in a range of activities across the primary curriculum. 151 Key Stage 1 and 2 pupils from four primary schools in the UK participated in the study. The findings revealed that understanding children's learning style preferences is an important consideration when deciding engagement methods for school projects.

2.6.2 Autism and learning style

Many pupils with ASD tend to be affected differently as a result of the same symptoms, so it is important to know how each child is impacted on in terms of their cognitive and interaction abilities. Hence, the specific requirements of each child,

which is known as learning style, should be identified (D'Elia et al., 2014). Pupils with ASD are more likely to rely on one style of learning, which could be determined by observing the person (Edelson, 2000). Many researchers (Banire et al., 2015; O'Connell et al., 2013; Snyder et al., 2012) found that many people with autism are visual learners as they can learn well through watching videos, pictures and movies. Another study (Lisle, 2007) found that Autistic people have learning preferences of kinaesthetic, visual, over auditory. (Roberts, 2010) found that people with autism who are kinaesthetic learners, learn better when “hands-on” activities with the teacher taught material are included to achieve learning. Nevertheless, research that has been conducted with autistic children regarding the impact of learning style is limited and needs further research. The research presented in this thesis will examine the VAK learning style with autistic and neurotypical children when integrated VAK into a conversational intelligence tutoring system.

2.6.3 Integrating learning styles into tutoring system

(Truong, 2016) reviewed the current application and integration of learning styles theories in tutoring system. (Truong, 2016) reviewed 51 studies and it is reported that there is a positive impact of tutoring systems using learning styles on student's satisfaction. In addition, a variety of statistical experiments have been identified, emphasising the positive effect of tutoring systems using learning styles on different factors such as performance, efficiency (time spent on different activities), cognitive loads and engagement. No known intelligent tutoring system using the VAK learning style model has been specific developed for autistic children.

2.7 Conversational Agents

Artificial intelligence is the field of developing intelligent machines and software that can learn, reason, gather knowledge, manipulate, communicate, and perceive the objects. Natural Language Processing (NLP) refers to Artificial intelligence (AI) methods of communicating with an intelligent system using a natural language (Pannu, 2015).

In 1950, a British researcher called Alan Turing published his famous AI article “Computing Machinery and Intelligence”. In this article, he discussed the possibility that one day machines can think like humans and he proposed the test known as (The Turing Test). In this test, a human interrogator interacts with a human and with a

machine and then he has to tell who is human and which one is machine. The computer passes the test if an interrogator fail to tell whether the written response is coming from the human or from the machine (Yaman, 2016).

After the Turing test, many AI researchers tried to achieve their dream of making the machine to communicate with human in natural language. Early attempts were computer programs known as chatterbots, that created the illusion of intelligence by applying tricks during the conversation, but in actual fact there is no intelligence programmed in to the system (Weizenbaum, 1966). The recent development of this field is Conversational Agents (CAs) that utilise artificial intelligence techniques to converse with the user. CA's provide communication between humans and computers using natural language. CAs have the ability to mimic human experts to offer advices and information to the user in specific domain (O'Shea et al., 2011).

A Conversational Agent (CA) is defined as 'a computer program which interacts with a user through natural language dialogue and provides some form of service'. (O'Shea et al., 2011). According to (Crockett et al., 2011b), the CA is defined as 'the ability to reason and pursue a course of action based on its interactions with humans and other agents'.

CAs have become a popular communication method and have been used in many application in different domain such as e-commerce, medication, education and so on (Campbell, 2016; Tegos et al., 2015). In addition, CAs are used effectively in many application such as student's debt management guidance (K. O'Shea et al., 2010), database interface (Owda et al., 2011), and as student's assistant by providing problem- solving advice as they learn (Alobaidi et al., 2013). Recently, many application of intelligent agents has been developed that focus on the virtual personal assistant, such as Microsoft's 'Cortana' (Microsoft, 2014), Apple Inc.'s "Siri" (Apple, 2014), and Google's 'OK Google' (Google, 2014). All these assistants are based on voice conversational agents; however the main functionality is similar to traditional CA approaches. The application synthesises the speech into text, then processes it by an engine to generate an appropriate response (Bellegarda, 2014).

There are two categories of CAs, 'Embodied CAs' (ECA) and linguistic CA's (LCA). Both categories are becoming increasingly popular in the field of CA. An embodied conversational agent (ECA) is defined as an agent which simulated face-to-face conversation with the user, by presenting the user with an avatar who talks to the user

and uses the naturalistic modes of communication such as hand, head, facial expressions and body gestures (Cassell, 2000). Linguistic conversational Agent (LCA) is defined as an agent who can converse with the user in written or spoken form (Yin et al., 2010)

2.7.1 Goal Orientated Conversational Agents (GO-CA)

The Goal Orientated Conversational Agents (GO-CA) is a specific family of CA that through the process of dialogue, GO-CA captures appropriate attributes to model the particular problem experienced by the user in order identify the appropriate solution (O'Shea et al., 2011). GO-CAs are designed to converse with humans through the use of natural language dialogue to achieve a specific task (Crockett et al., 2010). An example of GO-CA is the system developed by (Latham et al., 2010), which identifying and selling a person a mortgage or providing guidance through an organisation's policies and procedures in plain English. Pattern matching algorithms are used to capture the values of specific attributes through text based discussion with a user. The scripts which contain sets of rules about the domain are used and a knowledge base in order to guide the conversation towards achieving a specific goal. GO-CAs systems can provide consistent advice 24 hours a day in many different scenarios such as guiding a user through buying a suitable product, advising employees about their organisations policies and procedures, and tutoring a student to understand a learning objective (Crockett et al., 2011a). The majority of GO-CA's to date have been deployed in the English language. A small number of GO-CA's have been developed in Urdu Language and Arabic Language (Kaleem et al., 2014) (Alobaidi et al., 2013).

2.7.2 Conversational Agent Approach

There are three main approaches that are used to develop linguistic based CAs. Each will be briefly discussed. These are Natural Language Processing (NLP); Pattern Matching (PM); and sentence similarity measures.

2.7.2.1 Natural language processing (NLP)

NLP is defined as “the computational processing of textual materials in natural human languages” (Crystal, 2008). It was developed in the 1960s as a form of artificial intelligence (AI) research (Lester et al., 2004). Rapid developments in NLP have been

seen, such as development of large corpora of tagged texts (Penn Treebank)(Marcus et al., 1993), which accelerated the progress in NLP research (BNC-Consortium, 2003). Based on that progress, researchers were encouraged to develop CAs based on NLP, which focus on extracting utterances using grammar rules and list of attribute pairs from the conversation. This extracted information is used to fill a template-based response in order to generate a suitable response. Nevertheless, NLP based CA emphasis on structured lexicon, parsing and various other language handling aspects without reasoning about the meaning of the utterance. According to (Demir, 2003), an NLP-based CA should understand the user's utterance in order to produce human-like conversations. It means contain of expressed information should be recognized. There are certain other disadvantages of NLP-based CAs. First, the system might not recognise the misspelling of a word in the conversation and as a result the system may fail to understand the whole conversation. Second, NLP-based CAs may not comprehend a dialogue due to natural mistakes in grammar that people often make in everyday conversation. Thirdly, natural language is very ambiguous due to the richness in form and structure, which means that the word might have more than one meaning (lexical ambiguity) or a sentence might have more than one structure (syntactic ambiguity) (Fernezelyi et al., 2006).

2.7.2.2 Text-based Pattern Matching (PM)

PM is the process of matching a string or sequence of strings to a piece of text to find all occurrences of these strings inside that text (Hijjawi et al., 2014). From a CA perspective, algorithms are utilised within PM in order to handle user conversations by matching CA's patterns against a user's utterance. Wildcards, words and spaces generally compose the typical pattern. A wildcard is a symbol used to match a portion of the user's utterance (Kaleem et al., 2014). InfoChat (Michie and Sammut, 2001) as well as ALICE (Wallace, 2008) are two examples of CAs that have been incorporated effectively with PM algorithms. The typical procedure for such CAs is organizing a scripted domain into contexts, each of which has a number of related rules. Each rule has structured patterns that represent a user's utterance. PM is considered one of the successful methods for developing CAs (Bickmore and Giorgino, 2006). In contrast to possible alternative methods, PM is deemed to have various benefits. For example, PM has reduced operating and computing costs, as well as less time being required for

processing, due to there being no need for intricate procedures prior to processing. Furthermore, PM is effective regardless of language and is straightforward to comprehend. As a result, web-based interactions and conversations that are happening live and involve many individuals can be analysed and aided through CAs based on PM method (Allen et al., 2001). An additional advantage is that PM is able to overcome many linguistic challenges, which NLP approach faces, such as morphological changes occurring on words in the form of prefixes and suffixes. This is largely fixed through the utilisation of pattern wildcards that can amend mistakes in spelling and grammar, as well morphology amendments (Maragoudakis et al., 2001). The main limitation of using PM based CAs is the needs for a large number of patterns to implement a coherent domain. This limitation causes many issues. For example, certain keywords may have multiple meanings as well as singular and plural configurations, thus a programmer will have to take in to account these various possibilities and innate linguistic depth in terms of morphology. If one was dealing with Arabic language, then its various morphological amendments, prefixes and suffixes can be compensated for via utilising various methods, however the PM procedure can be particularly complex, especially in scripting. A further issue is that a user utterance may be written in multiple ways through adopting various alternative words and phrase structure.

2.7.2.3 Sentence similarity

Information, philosophy and linguistic theories have all been adopted as a means of assessing sentence similarity (Hatzivassiloglou et al., 1999). Generally, similarity as a term is used to describe the similarity level between two objects. In this section, three similarity approaches are discussed, which are: Lexical similarity, Semantic similarity and Hybrid similarity.

- a) **Lexical-based similarity approaches (LBS):** this approach provides the similarity on the basis of character matching (Rensch, 1992). It is measuring the distance between two strings to identify the similarity between them (Gomaa and Fahmy, 2013). LBS measures are divided into two categories: character-based and statement-based distance measures.

The character-based Similarity is based on the distance between each character in two strings such as Levenshtein distance similarity (Elmagarmid et al., 2007). For

Arabic language, many researchers used Levenshtein distance algorithm such as (Shaalan et al., 2012) who used it to develop Spell Checking tool for Arabic words. However, Levenshtein does not give accurate results when applied on the Arabic language. For example, when comparing (احمد) and (أحمد), the Levenshtein distance is equal to one, which means there is one different between the two words. The Arabic reader does not differentiate between (Alif with Hamza) and (Alif without Hamza) and therefore the distance should be zero (Abdel Ghafour et al., 2011).

The statement-based Similarity is based on the distance between each word in two strings such as Cosine similarity (Qian et al., 2004). For Arabic language, (Al-Ramahi and Mustafa, 2012) used cosine similarity to develop a system that determines the similarity between course descriptions of the same subject for credit transfer among various universities or similar academic programs. The results indicated that this technique has demonstrated better performance compared with other techniques such as Dice's Similarity. Another work in (Hajeer, 2012) studied the different lexical similarity measures in Arabic information retrieval and the results showed that the Cosine Similarity Measure is the best of other seven measures (Inner Product, Dice Coefficient, Jaccard Coefficient, Inclusion Similarity Coefficient, Overlap Coefficient Measure, Euclidean distance Measure and Manhattan Distance Measure).

- b) Semantic based similarity approach:** Semantics is “the study of the meaning of linguistic expressions”(Habash, 2010). Information retrieval, question answering, information extraction machine translation, data mining machine translation, and conversational agents have all been aided by means of similarity-based research, while it also has wider application to AI and computer processing as a whole. Applying the sentence similarity technique in CA developing is more effective than other techniques because it replaces the scripted patterns by a few natural language sentences in each rule. Consequently, the scripting effort is reduced to a minimum. Thus, a big reduction is caused in scripts and in the maintenance of such scripts (Li et al., 2004; O'Shea et al., 2008). Semantic similarity uses two main approaches to compute the similarity between strings: corpus-based and knowledge-based measures.

Corpus-based similarity (Islam and Inkpen, 2008) determines the similarity between words by using information from a large corpus, that is, a large body of different forms of text for comparison purposes.

Arabic is a poorly resourced language for corpus-based approaches in comparison with English, since there is a lack of sufficient data, and this negatively affects research into corpus-based and natural language processing in Arabic. More recently, a number of Arabic corpora have been developed; however, little overall improvement in the overall situation has been made (Alqahtani and Atwell, 2015). (Al-Thubaity, 2015) reviewed fourteen Arabic corpora and categorised them by their target language, designated purpose, text date, location, text domain, representativeness, mode of text, size, text type/medium, and balance. The paper reported some limitations of this review. Several of these corpora do not provide any information regarding the time period covered by the texts. In addition, the size of the reviewed corpora is large, such as the arabiCorpus (Parkinson, 2012), the arTenTen12 corpus (Belinkov et al., 2013), and the Leeds Internet Arabic Corpora (Sharoff, 2006). Moreover, for all corpora the component texts are not classified with respect to their dates or the time period to which they belong; there is therefore a limitation on the usability of the corpus and a difficulty in comparing languages used in different periods, and in observing how the Arabic language has evolved. For example, texts collected more recently are more likely to be written in Modern Arabic language, while texts collected from the old period are more likely to be written in classical Arabic language. Consequently, the efficiency of corpus-based similarity measurements will be affected because of the different texts language types. The design criteria for many of the corpora are unclear, which makes it difficult to evaluate accurately the findings of any research based on them, and there is no justification for excluding such criteria. A well-known early measurement technique, which utilise a corpus is, latent semantic similarity (LSA) (Landauer et al., 1998). It is defined as the similarity in meaning between two words. A matrix is formed in the LSA measure in which the row represents a particular word and the column represents the document or paragraph. To reduce the size of this matrix, the singular value matrix (SVM) approach could be used. After forming the matrix, the cosine similarity or Web Jaccard similarity algorithms could be applied to obtain the similarity between the words. LSA similarity is relatively fast in

comparison to other dimensionality reduction models (Tomás et al., 2006). The main disadvantage of this measure is that the similarity does not use any syntactic information from the compared texts. For example, the sentences “The dog hunted the man” and “The man hunted the dog” will be considered as identical (Al-marsoomi, 2015). (Reafat et al., 2012) developed an automatic Arabic essay scoring system using LSA similarity measurement. They improved LSA by processing the entered text before applying the measure. The new algorithm unifies the form of letters, deletes the formatting, replaced synonyms, and deletes "Stop Words" in order to create a matrix that represents texts more efficient than the general form of LSA matrix. The result indicates that the correlation between the human assessor and the system is 0.91, which is good in comparison with the human assessor and the traditional LSA that has a correlation varied from 0.78 to 0.87. In addition, (Al-marsoomi, 2015) developed a novel framework of an Arabic Short Text Semantic Similarity (STSS) measure, namely that of NasTa. The algorithm calculates the STSS based on Part of Speech (POS), Arabic Word Sense Disambiguation (WSD), semantic nets and corpus statistics.

Knowledge-based similarity (Budanitsky and Hirst, 2001) is a semantic-based similarity which determines the similarity between words using information derived from semantic networks. The most commonly-used semantic networks are WordNet (Miller, 1995) and the Natural Language Toolkit (NLTK) (Loper and Bird, 2002). WordNet is a database that collects words into sets of synonyms called synsets; for each synset, WordNet provides short definitions and usage examples, and records a number of relations, which include hyponyms, hypernyms, coordinate terms, holonyms and meronyms (Miller et al., 1990).

WordNet is used in various areas such as information retrieval (Kara et al., 2012), semantic similarity (Batet et al., 2011), and word sense disambiguation (Gutiérrez et al., 2010). Due to the success of WordNet in English language applications, several projects are currently being conducted to develop WordNet for other languages (Al-Zoghby et al., 2013). The Global WordNet Organisation, a non-profit public group, was established to support the development of WordNet and the production of dictionaries for other languages based on the English version of WordNet. It provides connectivity between the WordNet dictionaries in various

languages such as Arabic, Persian, Hebrew, Albanian and the African and Indian languages (GlobalWordNet, 2018).

Arabic WordNet (AWN) has been developed using the same methodology as EuroWordNet. It consists of 11,270 synsets and contains 23,496 Arabic expressions (words and multiwords) (Alkhalifa and Rodríguez, 2010). Numerous researchers have used AWN in their work, for example within an Arabic Information Retrieval system (Abderrahim et al., 2013) and an Arabic Question Answering (Q/A) system (AlAgha and Abu-Taha, 2015). The major limitations of the current AWN are a lack of information and concepts in comparison with WordNet in the English language, and some semantic relations between synsets. Many Arabic concepts have not been included in the AWN database. This limitation forms a major impediment for using AWN as the source for knowledge-based approaches. AWN could be enhanced and extended through several different approaches, for example the addition of new synsets, improvements to the existing synsets, improvements to the hyponymy/hypernymy relations, and the use of named entity (NE) synsets, ‘glosses’ and verb categorisations (Abouenour et al., 2010).

Several researchers have attempted to extend AWN through the addition of named entity synsets in order to enhance the performance of AWN. Al Khalifa and Rodriguez (Alkhalifa and Rodríguez, 2010) used Arabic Wikipedia to extract Arabic Named Entities (NEs) automatically and attach them to the Arabic WordNet. (Abouenour et al., 2010) focused on automated AWN improvement for the NE synsets, by using AWN in Q/A systems to improve system performance. However, these researchers reported that the coverage, hierarchy and relations associated with the NE synsets within AWN require improvement. They therefore proposed an approach based on the Yago ontology, a freely available ontology with a large coverage of NE. Their approach is based on the addition of both new NE and their super-types. In addition, (Saif et al., 2015) have mapped an Arabic WordNet synset to its corresponding article in Wikipedia, using monolingual and bilingual features to overcome the lack of semantic information in Arabic WordNet. This mapped method increased the coverage of Arabic WordNet by inserting new synsets from Wikipedia. Apart from this work focusing on NE synsets, the coverage of semantic relations is rather limited and needs to be developed.

A few researchers have attempted to overcome this limitation, such as (Boudabous et al., 2013) who suggested improving the performance of AWN using a linguistic method based on morpho-lexical patterns in order to add semantic relations between synsets using Arabic Wikipedia. However, the size of the Arabic Wikipedia remains an order of magnitude smaller than the English version, while the structured data in the Arabic version of Wikipedia, such as information boxes, are of lower quality on average in terms of coverage and accuracy. Consequently, the semantic similarity area using AWN needs further research in order to be more reliable. It is not sufficient enough to rely on AWN to apply semantic similarity measurements and to find the similarity between texts.

- c) **Hybrid based similarity approach:** This approach uses multiple similarity measures, and have been the subject of much research. (Mihalcea et al., 2006) tested eight semantic similarity measures, of which six were knowledge-based and two were corpus-based measures. These researchers evaluated the eight algorithms separately, and then combined them together, and found that the most efficient performance was achieved when several similarity metrics were combined into one. (Kaleem et al., 2014) used a hybrid approach which combined a lexical sentence similarity measure with pattern-matching techniques in order to calculate the matching strength between the user's expressions and the scripted patterns in Urdu language. The similarity algorithm comprises of Levenshtein algorithm to calculate the similarity between two strings, while Bipartite Matching (Dasgupta et al., 2008) is used to determine the word order variation and Kuhn-Munkres algorithm (Burkard and Cela, 1999) is used to find the maximum sum of a given matrix of weights. This hybrid algorithm overcomes the challenges of string similarity and free word order in Urdu language or other languages with free word order such as Arabic language. In Arabic language, (Gomaa and Fahmy, 2012) used lexical-based and corpus-based similarity measures to develop their Short Answer Grading system. They tested the measurements using three stages. In the first stage, they assessed the similarity between the model answer and student answer using thirteen lexical-based measurements, of which seven were statement-based and six were character-based measures. Their results showed that the best correlation values achieved using character-based and statement-based measures were obtained using the n-gram and block distance (Manhattan distance) approaches respectively. In the

second stage, they measured the similarity using corpus-based similarity measurements (P. Kolb, 2008): DISCO1 (Computes the similarity of the first order between two words based on their collocation sets) and DISCO2 (Computes similarity of the second order between two words based on their sets of distributing similar words). The results showed that DISCO1 achieved more efficient correlation values. In the third stage, the similarity was evaluated by combining lexical-based and corpus-based measures. The best correlation value was obtained from mixing n-gram with DISCO1 similarity techniques. From these researches, the hybrid similarity approach is considered as a promising approach with Arabic language because it combines more than one type of measurements which leads to the similarity being more robust.

2.7.3 Conversational Agents in Arabic Language

The Arabic language is an official language of more than twenty countries, and a major spoken language by over 300 million people worldwide (Shalan, 2010). The ability to develop an Arabic CA system is a significant challenge. The Arabic language is considered as being a complex language, which is often ambiguous in nature.

Some researchers developed CA in an Arabic language, the first example is (Shawar and Atwell, 2004), who developed an Arabic Chatbot, which accepts the user utterance related to Islamic issue and then answers the questions from the Quran. The limitation in this system that it generate an answer and send it to the user even if the answer is not related to the question.

Another example is ArabChat, which is developed by (Hijjawi et al., 2016). It is designed for Science University in Jordan as an information point advisor for their native Arabic students. It is based on Pattern Matching approach to handle and modelled based upon the Infochat (Michie and Sammut, 2001). The main limitation in ArabChat is that it is depend on the pattern matching technique with a vast amount of scripted patterns. These patterns must cover all expected keyword's changes to meet different cases of the scripted keyword such as singular/plural. In addition, using this amount of patterns make it difficult to maintenance the system and time-consuming process. (Noori et al., 2014) developed an Arabic CA architecture based on a pattern matching algorithm for the development of a goal orientated Arabic Conversational Agents (ACA). It is developed for the Iraqi passport system. The main limitation in

this system is using Pattern Matching technique only, which causes domain limitation, and the complexity associated with scripting and maintaining.

2.7.4 Challenges within the Arabic Language

The research into CA's has been focused on mainly English and western languages. For Arabic language, CA research is still in its early stages and do not have the extensive lexical infrastructures that are required such as Corpora and WordNet. PM remains the predominant methodology for scripting, as other methodologies require Corpora or WordNets which are still not readily available in Arabic languages. Arabic language has many features, which are considered as challenges for Arabic CA. The first challenge is that there are three types of Arabic language, which are known as Classical, Modern and Colloquial. Classical Arabic, which is used in the Quran, is more complex in its grammar and vocabulary than modern Arabic. It has a large number of diacritics that facilitate word pronunciation and detection in their grammatical cases. The second type is the Modern Arabic, all diacritics were omitted to make the process of the reading and writing easier and fast. This type is considered as the official language of the Arab countries and is used in everyday language, education, and in the media. Usually, computational Arabic-based researches use Modern Arabic (Habash, 2010). In Colloquial Arabic, which is the third type, the grammar and vocabulary are less sophisticated in comparison to Modern Arabic. However, most people use it in their everyday spoken conversations and in informally written letters because of its simplicity (Al-Saidat and Al-Momani, 2010). Arabic people make many mistakes in grammar when they using Modern Arabic and they mix between Modern and Colloquial Arabic. Moreover, different Arab countries have different Colloquial Arabic. Consequently, the challenges increase for Arabic CA to recognise the user's utterance. The second challenge is the Arabic Morphology. Arabic language is complex because of the morphological variation and the agglutination phenomenon (Kadri and Benyamina, 1992). The form of letters change depending on their position in the word. In addition, the word may consist of prefixes, lemma and suffixes in different combinations, which results in a very complicated morphology (Abuleil et al., 2002). These features will increase the challenge for Arabic CA to determine the grammatical case for a user's utterance words. The third challenge is the capitalisation. Arabic language does not support capitalisation of

proper nouns such as countries names, people names. Whereas, in Latin languages these begin with a capital letter (Alanazi et al., 2015). Arabic CA may not recognise these words and this increases the difficulty of detecting such nouns in textual Arabic conversations. The fourth challenge is that the short vowels and diacritics are optional in Arabic. They are needed for pronunciation and disambiguation. However, Modern Arabic texts do not include diacritics. Therefore, the form of a word may have two or more different meanings depending on the context. This creating an ambiguity during the conversation in Arabic CA. These challenges make it difficult to use existing conversational agent engines, scripting methodologies, algorithms and approaches, developed for English language, and apply it for Arabic text. A new architecture needed to support conversation tutoring system in the Arabic language, which can incorporate the issues highlighted efficiently to produce a functional Arabic CITS for Autism.

2.7.5 Conversational Agent Knowledge Base Development

The development of a Knowledge Base System (KBS) is the process of transferring the human knowledge into an implemented knowledge base. Usually, the required knowledge is obtained by interviewing experts on how they solve specific tasks (Musen, 1993). This knowledge is implemented as production rules that are then executed by an associated rule interpreter. The initial phase of creating KBS is knowledge extraction, where the required knowledge is collected and stored (Morik, 1991). It is reported (O'Shea et al., 2011) that knowledge usually extracts information about a domain from many different sources, including:

- Who will use the CA being developed.
- Managers in the client organisation.
- Documented procedures of the client organisation (e.g. workflow charts).
- Telephone logs of customer calls related to the domain.
- 3rd party websites (e.g. government legislation concerning the domain).
- Practitioners in the client organisation who interact with the customers.

2.7.6 Conversational Agent Evaluation

Due to the absence of established evaluation methodologies for assessing conversational system, it is rather problematic to undertake the process. However, as a means of evaluating CA, certain researchers (Alobaidi et al., 2013; J. O'Shea et al., 2011; Kaleem et al., 2014) have identified relevant factors such as system performance and user satisfaction. (Schrepp et al., 2013) proposed a framework through User Evaluation Questionnaires (UEQ) to measure software quality and user experience and in interactive software. Based on this framework, the evaluation of software quality is divided into two categories, “pragmatic quality” and “hedonic quality”. Pragmatic quality is related to task quality such as the effectiveness and efficiency of the task completion. Hedonic quality is related to non-task aspects such as user stimulation and aesthetic impressions. These two categories can be translated into objective measures and subjective measures respectively.

Both objective and subjective measures can be utilised in order to evaluate CA/ Dialogue system success and efficacy, in accordance with the recommendations of various researchers over a long period of study. Using both of objective and subjective evaluation measurements ensure that both of the effectiveness of the CA's functionality and the usability from the user perspective are tested. The evaluation of LANA will consist of a combination of subjective and objective measures, based on similar methods used in the literature as well as methods adapted to suit LANA and the research questions (see chapter 4 for LANA CITS evaluation method).

In subjective evaluation, (Hone and Graham, 2000), (Silvervarg and Jönsson, 2011) and (Rauschenberger et al., 2013) utilised a questionnaires method in order to determine subjective characteristics such as user satisfaction. The ratings are applied either, by users during or right after the dialogue or by experts by analysing recorded dialogues. (Ultes et al., 2013) argued that asking real users interacting with the system are the most suitable individuals for evaluating user Satisfaction. Therefore, it is proposed that the most efficient and effective method to evaluate the end user satisfaction is to conduct a questionnaire to the participants to obtain their individual opinions with regards to the system usability.

Satisfaction from the users' perspective can be determined via a questionnaire. The questionnaire can be utilised to highlight the user's opinion on the following subjective attributes:

- Usability
 - Ease of use
 - Interface design
 - Language used
- Performance/User satisfaction
 - Was the goal/task achieved (Pietquin and Hastie, 2013)
 - Was the information given understandable
 - Time taken to reach aim/goal
 - Was the CA approachable/intuitive
 - CA naturalness (Lutfi et al., 2013)
 - Domain knowledge coverage

Objective measures are typically used for analysis in the majority of research (O'Shea et al., 2011), however there is often a discrepancy between subjective measures outlined at the outset of research and objective measures. (Walker et al., 1997) employed the PARADISE model as one of the rare methodical and logical processes for analysis, which since the time of their paper publication has influenced the approach of much other research. The work in this thesis also utilises PARADISE alongside various context specific procedures. Objective measures can be measured through records and logs of the user's dialogue with the CA. The logs are used to capture and store variables, which are related to the dialogue such as user utterance, CA response, rule fired, similarity strength etc. the following attributes can be measured and analysed using this information:

- Dialogue / Conversation length
- Conversation success and goal achievement
- Effectiveness of the similarity algorithm in combination with wild card PM.
 - ability to reduce the number of scripted patterns and scripting time
 - effectiveness to calculate similarity strength
 - utterance recognition accuracy measures

2.8 Intelligent Tutoring Systems

Intelligent Tutoring Systems (ITS) include any computer program used in learning and contains intelligent technologies to personalise learning according to individual

student characteristics, such as knowledge of the subject, emotion, and learning style (Dermeval et al., 2017). Many recent positive reviews support the effectiveness of ITSs (Kulik and Fletcher, 2016; Crow et al., 2018). Thus, in many common academic subjects, ITS can successfully complement and substitute other instructional models at all educational levels (Ma, 2017).

2.8.1 Architecture of ITS System

An ITS system typically, has four basic components: the domain model, the tutoring model, the student model, and the user interface model (Ahuja and Sille, 2013).

- The domain model, which known as the cognitive model or expert knowledge model, emphasizes all facts, rules, concepts, and problem-solving strategies of the domain in context. It is considered as a source of knowledge, a diagnostic of errors, and a standard for evaluation of the student's performance.
- The student model contains information about individual student characteristics (e.g. learning style), which can be used to adapt the system to the student. Therefore, student model is a vital tool for the adaptation process.
- The tutoring model, which is also called the teaching strategy, receives information from the domain models, student models and devises tutoring strategies with actions such as course preparation and adaptation. This model organizes instructional interactions with each student. It is linked to the student model to structure the knowledge model based on the student's characteristics (e.g. learning style).
- The User Interface Model, is composed of such elements as the keyboard, mouse, word, image and various multi-media aspects necessary for the user to interrelate to the ITS (Padayachee, 2002). The prime factors for user-acceptance are user-friendliness and presentation. The Figure 2.1 presents a typical ITS architecture.

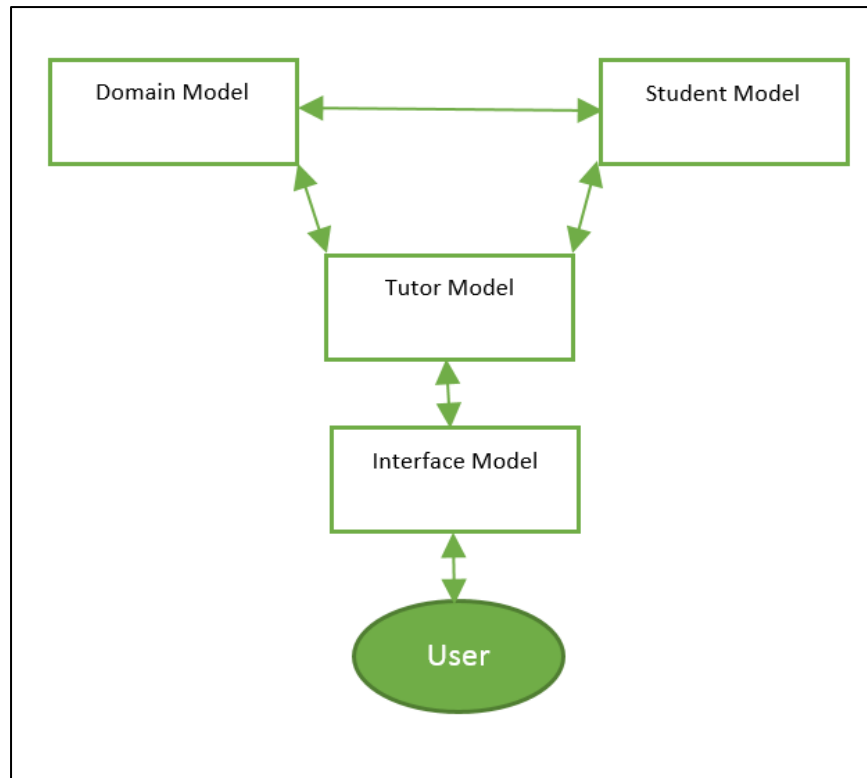


Figure 2.1: ITS architecture (Ahuja and Sille, 2013)

2.8.2 Conversational Intelligent Tutoring Systems

Conversational Intelligent Tutoring Systems (CITS) are an extension of ITS. It integrates natural language interfaces rather than menus to allow users to learn topics through discussion as they would in the classroom. Many researchers have and continue to, develop CITS for different domains. AutoTutor (Graesser et al., 2005) is one examples of CITS, which is a computer tutor that is designed to assist students in the university level to learn an introductory computer literacy course. The system teaches the fundamentals of hardware, operating systems, and the Internet. However, AutoTutor does not consider the student learning styles during a tutoring session. It means that all the students are provided with the same learning experience regardless the ability. Oscar (Latham et al., 2014; Crockett et al., 2017) is another example of CITS, which overcomes AutoTutor limitation. Oscar leads a tutoring conversation and it dynamically predicts and adapts to a student's learning style. Oscar is developed for students in University's level and teaches SQL course. In the Arabic language, CITS is considered as a new area of research. Recently, (Alobaidi et al., 2013) developed an Arabic CITS (Abdullah) for modern Islamic education. Abdullah CITS is an online

system that aims to teach children aged 10 to 12 years old the essential topics in Islam. The system allows conversation, discussion and interpretation with verses in classical Arabic language by engaging in dialogue using Modern Arabic language. To our knowledge, no academic research exists on the Arabic conversational intelligent tutoring system for Autistic children.

2.8.3 Evaluation of CITS

According to (Graesser et al., 2005), who developed AutoTutor, measuring learning gain of the students is considered the effective methodology. The learning gain in AutoTutor have been evaluated based on measuring the student's understanding at the end of the tutorial. This is also supported by recent work of (Latham et al., 2014; Crockett et al., 2017). The learning gain measured by providing a pre-test (before the tutorial) and post-test (after the tutorial) and then compare the student's score test (Kelly and Tangney, 2006; Graesser et al., 2005; Lee et al., 2006). Relative learning gain was also measured. Relative learning gain is a measure that calculate the average improvement in test scores as a percentage of the possible improvement (Latham et al., 2014). This measure additionally takes into account the opportunity for improvement. For example, if student get 8/10 in pre-test and only improves by 1, this is different to another getting 3/10 in pre-test and only improving by 1 – improvement is 50% in first case but only 14.2% in the second case. The subjective evaluation of CITS is usually conducted using questionnaire in order to know the user's evaluation to the CITS.

2.9 Chapter Summary

It can be concluded from the literature review that technology-based intervention and the TEACCH method for pupils with ASD are promising strategies as educational intervention for treating them specially pupils with ASD who are included in mainstream schools. The traditional process of facilitating their learning and participation remains complex because of lacking specialised teachers to deal with their specific needs. In Addition, many pupils with ASD tend to be affected differently, so it is important to design specific requirements of each child, which is known as learning style. VAK theory is considered to be one of the classical learning theories in the educational field that classifies learners by sensory preferences: Visually (V), Auditory (A), and Kinaesthetic (K). No known intelligent tutoring system using the

VAK learning style model has been specific developed for autistic children. The aim in this research is to develop an Arabic Conversational Intelligent Tutoring System (CITS) that adapt to the learning styles for autistic children and enhance their learning. However, the research into CA's has been focused on mainly English and western languages. For Arabic language, CA research is still in its early stages and do not have the extensive lexical infrastructures that are required such as Corpora and WordNet. The use of similarity approaches within the Arabic language poses significant challenges.

1. Arabic is a complex language, which is often ambiguous in nature.
2. Currently, Arabic WordNet lacks the necessary information and a range of concepts in comparison with the English language WordNet, and some semantic relations between synsets are missing.
3. Moreover, most of the Arabic corpora does not cover all possible domains and words, with each corpus being focused on a specific domain and lacking information.

Using the lexical similarity approach within Arabic CAs is not reliable due to the different features in Arabic language such as Arabic morphological features. The semantic similarity approach is not reliable with Arabic CA because of the weakness of AWN and Arabic Corpora (Noori et al., 2018). In most recent times, the hybrid similarity approach is considered as a promising approach with Arabic CA because it combines more than one type of measurements which leads to the similarity being more robust. A new architecture needed to support conversation tutoring system in the Arabic language, which can incorporate the issues highlighted in this chapter efficiently to produce a functional Arabic CITS.

Chapter 3 – Research Methodology and Procedure

3.1 Overview

Chapter 2 provided a review of literature in four key areas related to the research presented in this thesis, including: educational interventions for children with ASD, conversational agent (CA) architectures with Arabic language, Conversational Intelligent Tutoring System (CITS) architectures with Arabic language, and the impact of learning style models for pupils with autism. This review of literature provided guidance in the development of an Arabic conversational intelligent tutoring system for children with ASD (LANA-I, described in Chapter 4).

This chapter details the approach used to develop and evaluate LANA CITS. It describes the research design, ethical consideration, data collection, and data analysis that were used in this research. In this chapter, section 3.2 describes the methods used in the development and evaluation. Section 3.3 illustrates the data collection and the analysis methods used. Section 3.4 discusses the ethical consideration associated with the data collection, and section 3.5 summarises the chapter.

3.2 Research Design

The research presented in this thesis has multiple objectives: First, research current education technology for children aged 10-12 who exhibit ASD in order to develop a novel education system that is not developed before. Second, undertake a review of CAs with Arabic Natural Language Processing resources for Arabic short text similarity and assess viability for incorporation in an Arabic CITS. Third, for a chosen domain (Science), design two educational tutoring scenarios, which enable capture of autistic learning styles and adaptation of the tutorial based upon the learning style model. Fourth, design and implement architecture for Arabic CITS for appropriate education scenario. Fifth, design an experimental methodology to validate the educational tutoring scenario in the Arabic CITS and conduct a study in Saudi Arabia to evaluate the ability of CITS to adapt to autistic children learning style.

3.2.1 Development of LANA CITS

LANA CITS is a Conversational Intelligent Tutoring System that uses the Visual, Auditory, and Kinaesthetic learning style (VAK). It supports learning in autistic pupils, who are studying in mainstream primary schools. Based on the research conducted into the development of Arabic CITS and the complexities of the language challenges (see Chapter 2), lack of Arabic CITS is due to the lack of Arabic linguistics resources. In addition, no Arabic CITS exists for Autistic children. Literature suggested that many pupils with ASD tend to be affected differently, so it is important to design specific requirements of each child, which is known as learning style. VAK theory is considered to be one of the classical learning theories in the educational field that classifies learners by sensory preferences: Visually (V), Auditory (A), and Kinaesthetic (K). Consequently, VAK learning style model was used in this research to adapt the tutorial to the autistic children learning style. In terms of CA, the research into CA's has been focused on mainly English and western languages. An Arabic CA research is still in its early stages because of the challenges of using similarity approaches within the Arabic language. For example, Arabic is a complex language, which is often ambiguous in nature. In addition, Arabic WordNet lacks the necessary information and a range of concepts in comparison with the English language WordNet. Moreover, most of the Arabic corpora does not cover all possible domains and words, with each corpus being focused on a specific domain and lacking information. Therefore, the semantic similarity approach is not reliable with Arabic CA because of the weakness of AWN and Arabic Corpora. The lexical similarity approach with Arabic CA is also not reliable due to the different features in Arabic language such as complex word structure, lack of capitalization, and minimal punctuation. From the review of literature, the hybrid similarity approach is considered as a promising approach with Arabic CA because it combines more than one type of measurements which leads to the similarity being more robust. Therefore, the hybrid similarity approach was used in this research to develop an Arabic CA. LANA CITS was developed through two prototypes, the first prototype LANA-I was developed with the following main features:

- Ability to personalise the tutorial with the user's learning style (VAK) and provide suitable material to the user according to the user's learning style

(images and videos for Visual learner, Sound for Auditory learner, physical models for Kinaesthetic learner).

- A novel CA engine that uses the pattern matching technique to find a suitable response to the user's utterance in the first instance. If no suitable response is found, the engine uses the short text similarity technique to find a response. Combining these techniques reduces the number of scripted patterns and therefore scripting effort.
- Managing the response when the context is changed. For example, creating the right response when the user writes something that is not related to the tutorial topic then the user is brought back to the point where the conversation is interrupted.

The second prototype is LANA-II which was developed to overcome the weaknesses of LANA-I:

- A new string similarity algorithm and pre-processing method have been developed and added to the architecture in order to improve the robustness and accuracy of LANA-II engine.
- TEACCH method has been added to the GUI in order to adapt the tutorial environment to the autistic students learning.
- The knowledge base is expanded by adding a new tutorial and expanding the general topics in order to address the unrecognised utterances from LANA-I.

Details of the system development are presented in Chapter 4 for the first prototype and Chapter 6 for the second prototype.

3.2.2 Evaluation of LANA CITS

Following its development, LANA CITS was evaluated through two phases. The first phase was to evaluate the first prototype LANA-I with neurotypical students to test two hypotheses and their questions:

Hypothesis A: LANA-I can be adapted to the student learning style:

Question1: Can LANA-I which embeds the learning Style VAK help a student to improve their learning?

Question2: Can LANA-I deliver personalised adaptation of a tutorial to the student's learning style successfully?

Hypothesis B: LANA-I is an effective Arabic CA:

Question 1: Are the students satisfied with the usability of LANA-I?

Question 2: Is the Arabic Conversational Agent used within LANA-I robust?

The participants were on the target age group (10-12) years old whose first language is Arabic. None of the children has been diagnosed as Autistic. It is important to test the quality of the tutorial and the materials, which are used in LANA-I, with the general population before testing it with Autistic children. This is because the system robustness and its ability to complete the task must be tested with the general population before autistic children in order to avoid any confusion that may happen to the autistic child during the tutorial. Subjective and objective evaluation methodologies were used to evaluate the first prototype (Chapter 5).

The second evaluation phase was to evaluate the second prototype LANA-II using three experiments. The first experiment was conducted to test the following hypothesis and its questions:

Hypothesis A: LANA-II can be adapted to the autistic students learning style:

Question1: Can LANA-II improve the autistic students' perceptions of the learning experience by adapting to VAK learning Style?

Question2: Can LANA-II by adapting to VAK learning style improve the learning gain of the autistic students?

The participants were autistic children on the target age group (10-12) years old whose first language is Arabic and have a high functioning autism (HFA) who they have no problem with language and intellectual disabilities. This experiment was conducted to measure the capabilities from different aspects such as tutoring success, VAK learning style model and user evaluation. The second experiment was conducted to test the following hypothesis and answer its questions:

Hypothesis B: the enhancements made to LANA-II architecture improve the overall effectiveness of LANA-II engine:

Question 1: Is LANA-II engine more effective than (LANA-I) engine?

Question 2: Do the improvements added in the LANA-II engine improve perceptions over (LANA-I) by the user?

The participants were on the target age group (10-12) years old whose first language is Arabic. None of the children have been diagnosed as Autistic. This experiment

aimed to test the enhancement made in LANA-II engine to compare the results from the first prototype and the results from the second prototype system.

The third experiment was a qualitative experiment (case study) that contributes to test the following hypothesis and answer its question:

Hypothesis C: LANA-II is appropriate as a virtual tutor for Autistic students:

Question: Are the autistic students satisfied with the usability of LANA-II?

The case study with three autistic students to test if LANA-II is appropriate as a virtual tutor for Autistic students. The results derived from the three experiments will contribute towards concluding the main research question. Details of the system evaluation methods and results are presented in Chapter 5 for the first prototype and Chapter 7 for the second prototype.

3.3 Data collection and analysis methodology of the research

3.3.1 Triangulation

Triangulation is often used to describe research where two or more methods are used, known as mixed methods. Combining both quantitative and qualitative methods to answer a specific research question may result in one of the following three outcomes: (1) the results may converge and lead to the same conclusions; (2) the results may relate to different objects or phenomena but may be complementary to each other and used to supplement the individual results and (3) the results may be divergent or contradictory. Converging results aim to increase the validity through verification; complementary results highlight different aspects of the phenomenon or illustrate different phenomenon and divergent findings can lead to new and better explanations for the phenomenon under investigation (Creswell et al., 2017).

This research study employed Triangulation method by using both quantitative and qualitative methods in order to answer the main research questions (Chapter 1, section 1.3). It uses a combination of qualitative and quantitative methods to highlight the strengths and perspectives of each method. It also provides additional evidence and support for the findings in order to answer the main research questions. The quantitative methodology includes data from the LANA log file (chapter 4, section 4.4.6) and the questionnaire that was conducted after the tutorial (chapter 5, section 5.2.2). The statistical Package for the Social Sciences (SPSS) was used for the statistical analyses for quantitative data. Required analysis was done with the aid of

SPSS software to get the results which were analysed. In the qualitative methodology, three case studies were conducted and the data was gathered from the interview and the questionnaire. This data was analysed using the following methodology:

1. Data management: This involves organizing and gathering all the data. In this study, the interview was transferred into a written format.
2. Reading and noting: This involves reading through all the transcriptions and making notes in the margins regarding the interviewee's comments.
3. Describing: this involve describing the participant's experience with the study.
4. Classifying: The researcher then made a list of significant statements, quotes that emerged from the data and then grouped them together.
5. Interpreting: the researcher developed textural and structural descriptions of the data. Textural descriptions are the written descriptions of what the participants experienced. After the textural descriptions are written, the researcher outlined the contextual influence, and this is the structural description.
6. Representing and visualizing: After both the textual and structural descriptions are written, the researcher wrote the findings of the study. In addition, tables were created to help illustrate the data.

3.4 Ethical consideration

The current study was subject to certain ethical issues because the participants were children. MMU ethical approval was obtained (No. SE151615, Appendix L). An informed consent form (see Appendix E) and information sheet (see Appendix D) were sent to the participants' parents and the School. The information sheet outlined the details and purpose of the study, and the benefits for participating in the study. They also informed that the information gathered from the participant is stored securely and none of children's real names will be recorded.

3.5 Chapter summary

This chapter presents the methods used to develop an Arabic conversational intelligent tutoring system for children with ASD. It describes the research design, ethical consideration, data collection, and data analysis that were used in this research.

The methods to enable development of LANA CITS were conducted through two steps process. The first is to design and adapt the tutorial to the children learning style using the VAK model. The Second process was to develop the Arabic CA using the hybrid similarity measures. After designing and implementing an Arabic CA, the Arabic ITS interface was integrated with the CA.

To evaluate the system, two methods of data collection and analysis were used: First, the quantitative method, which includes data from the LANA log file and the questionnaire that was conducted after the tutorial. This data was analysed using SPSS software. Second, the qualitative method, which includes three case studies and the data was gathered from the interview and the questionnaire.

Details of the development and evaluation for the first prototype is described in Chapter 4 and Chapter 5. The second prototype development and evaluation were presented in Chapter 6 and chapter 7.

Chapter 4 - Methodology for Developing the Arabic CITS (LANA-I)

4.1 Overview

This chapter describes the architecture and components developed for implementing an Arabic Conversational Intelligent Tutoring System for Autism called LANA-I. LANA-I is a novel Arabic CITS, which delivers topics related to science subjects by engaging with the user in the Arabic language. LANA-I uses Visual, Auditory, and Kinaesthetic (VAK) learning style model to personalise the lesson according to the user learning style preference.

An intelligent computer tutor could have a positive impact on distance learning and offer support for traditional class-based courses especially for Autistic people who have a difficulty in the social and communication interactions. Autistic children taught by computer tutors gain more information than those taught by human tutors (see chapter 2 section 2.5.2). However, based on the research conducted into the development of Arabic CITS and the complexities of the language challenges (see Chapter 2), little works have been done on methods for developing Arabic CITS mainly due to the lack of Arabic linguistic resources. In addition, no work has tackled the development of Arabic CITS for children with autism especially high functioning autism or Asperger syndrome.

This chapter describes the methodology of developing LANA-I, which includes an architecture for an Arabic CITS, a mapping of a tutorial on science subjects to the VAK learning style model, development of an Arabic CA, and a new Arabic scripting language combining pattern matching (PM) and short text similarity (STS).

In summary, the main features of LANA-I are:

- Ability to control and manage the conversation through context. (Context is defined in section 4.1).
- Ability to personalise the tutorial with the user's learning style (VAK) and provide suitable material to the user according to the user's learning style (images and videos for Visual learner, Sound for Auditory learner, physical models for Kinaesthetic learner).

- A CA scripting language, which provides Arabic dialogue for LANA-I and allows the capture of variables from the user utterance.
- A novel CA engine that uses the pattern matching technique to find a suitable response to the user's utterance in the first instance. If no suitable response is found, the engine uses the short text similarity technique to find a response. Combining these techniques reduces the number of scripted patterns and therefore scripting effort.
- Managing the response when the context is changed. For example, creating the right response when the user writes something that is not related to the tutorial topic then the user is brought back to the point where the conversation is interrupted.

In this chapter, section 4.2 outlines the LANA-I architecture. Section 4.3 details the methodology of implementing LANA-I. Section 4.4 illustrate the implementation of the first component of LANA-I, which is the Knowledge Base, Section 4.5 describes the implementation of the second component of LANA-I, which is the Arabic conversational agent. Section 4.6 details the implementation of the third component of LANA-I, which is the intelligent tutoring system. Section 4.7 discusses the LANA-I Workflow and section 4.8 summarises the chapter.

4.2 LANA-I Architecture Overview

This section illustrates the design and development of the LANA-I architecture. LANA-I was developed based on two phases. The first phase is designing and implementing an Arabic CA, and the second phase is integrating the Arabic ITS interface with the CA.

4.2.1 Phase 1: Designing an Arabic CA architecture

The Arabic CA was developed based on a number of features, which can be summarised as follows:

- Capturing variables from the user utterance.
- Ability to control the conversation through context.
- Ability to personalise the lesson with the user's learning style and provide suitable supporting material to the user.
- A scripting language to provide Arabic dialogue for LANA-I.

- Managing the response when the context is changed.

The proposed framework for the Arabic CA consists of five components as shown in Figure 4.1:

1. **Graphical User Interface:** Responsible for the communication between the user and the agent through a web interface (section 4.6.1).
2. **Controller:** The controller manages the conversation between the user and the Arabic CA, as well as cleaning the utterance by removing diacritics and other special characters (e.g. ! £ \$) (section 4.5.1).
3. **Conversational Agent Manager:** Responsible for controlling and directing the conversation through contexts. In addition, the manager must ensure that the discussion is directed towards the end goal of the tutorial (section 4.5.2).
4. **Conversational Agent engine:** Responsible for pattern matching and calculating the similarity strength between the user utterance and the scripted patterns to generate the response (section 4.5.3).
5. **The knowledge base:** contains conversation log file, and the domain knowledge in a relational database which includes: CA scripts, General contexts such as weather, agreement and rude words (section 4.4).

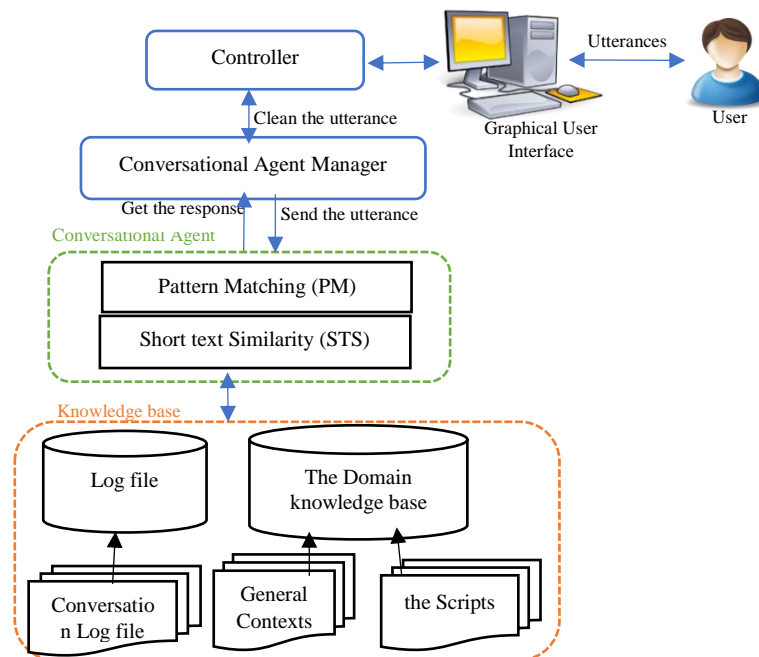


Figure 4.1: The Arabic CA architecture

4.2.2 Phase 2: Designing the ITS architecture

In this phase the ITS architecture is designed to adapt to the user's learning style with the Arabic CA. Based on the typical ITS architecture, which was described in chapter 2 section 2.8.1, ITS architecture in LANA-I consists of four main components as shown in Figure 4.2:

1. **User Interface Model:** responsible for the interaction between the user and the ITS components (section 4.6.1).
2. **Tutor Model:** the ITS manager, which is considered as the main component that interacts with the user through the GUI, and personalises the tutorial based on the user's learning style (section 4.6.2).
3. **Student Model:** a user's profile, which records the user's responses during the tutoring session and record the student's profile such as user ID, user's age, gender, user's learning style, and pre-test and post-test scores (section 4.4.6).
4. **Domain Model:** The Domain Knowledge Base, which contains structured topics to present to the user (section 4.4).

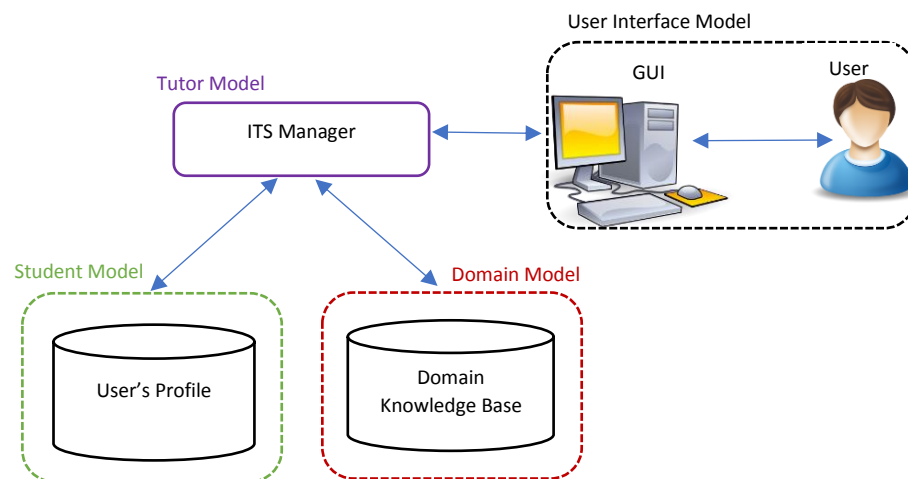


Figure 4.2: LANA-I ITS Architecture

4.3 Methodology of Implementing LANA CITS

The LANA-I was designed and implemented according to the GO-CA development methodology (Crockett et al., 2011b) (Kaleem et al., 2016). The development methodology combines elements of the staged approach used in the Waterfall model with elements of prototyping development. Figure 4.3 shows the major stages.

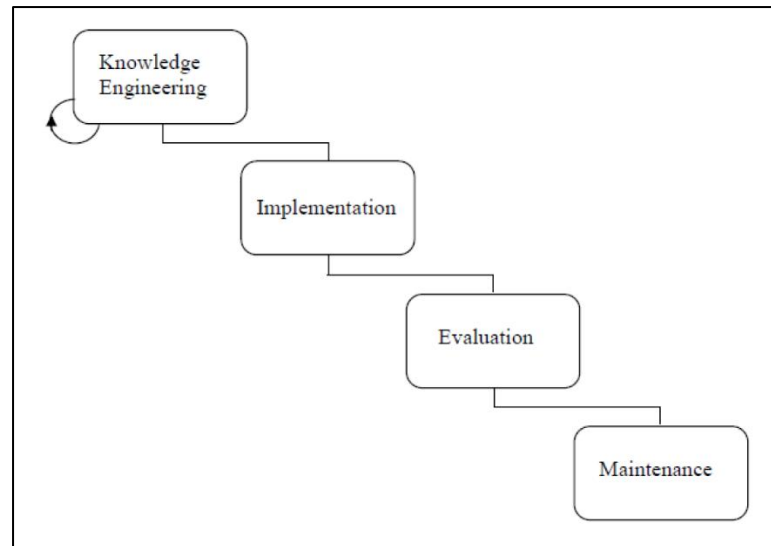


Figure 4.3: GO-CA Software Development Methodology (Kaleem et al., 2014)

LANA-I architecture was designed by combining the Arabic CA (Figure 4.1) and ITS architecture (Figure 4.2). The proposed LANA-I architecture has three components as shown in Figure 4.4.

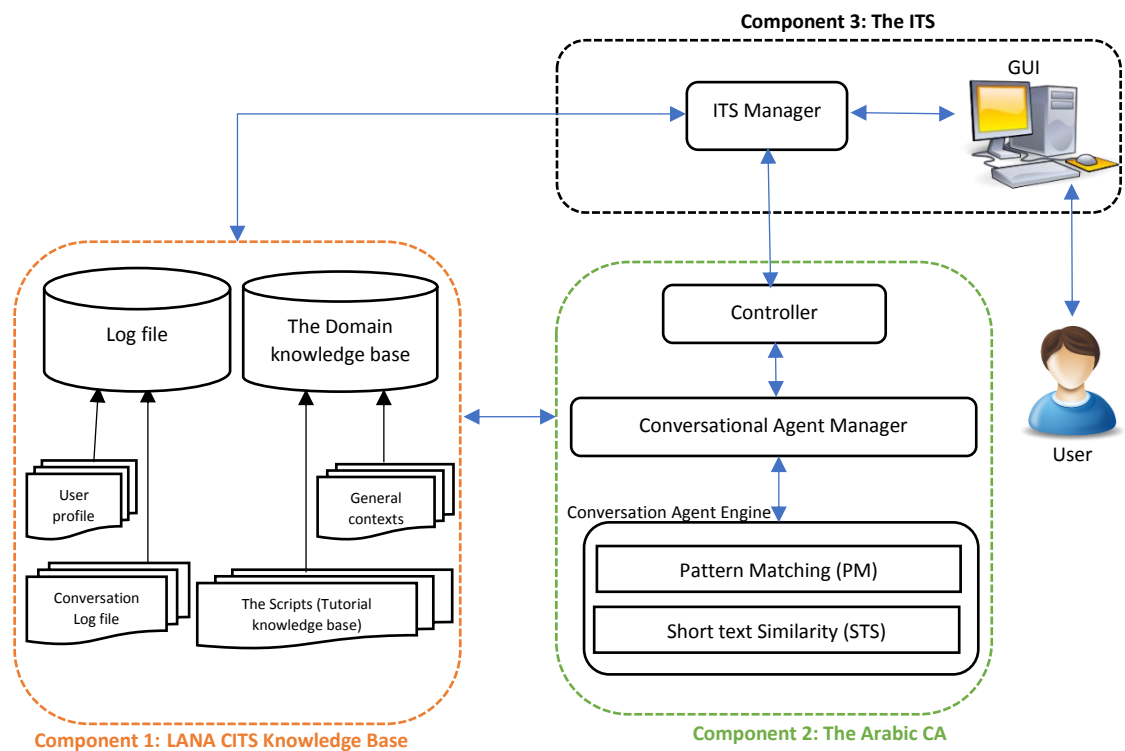


Figure 4.4: LANA-I architecture

These are: component 1 (LANA-I Knowledge Base), component 2 (The Arabic CA), and component 3 (the ITS). The following sections describe the implementation for each component in detail

4.4 Component 1: Implementing the LANA-I Knowledge Base

The first component consists of:

1. The domain Knowledge base, which includes:
 - a. Arabic general contexts such as weather, and greeting.
 - b. The Scripts (Tutorial knowledge base).
2. The log file, which includes:
 - a. User's profile.
 - b. Conversation log file.

4.4.1 The Domain

The domain used to script the CA for LANA-I was from the Science curriculum for the age (10-12 years old). The curriculum was provided by the Ministry of Education in Saudi Arabia (Education, 2017). The scripts consist of a context structured according to the topics in the science book such as the Earth, Moon, Solar System, eclipse, etc.). LANA-I is therefore designed to deliver the Science topic to Autistic learners in the Arabic language based on the science book (Education, 2017) that is taught in schools in Saudi Arabia.

4.4.2 Arabic CITS knowledge engineering

Based on (Jee et al., 2014) definition, knowledge engineering is the process of extracting, representing, encoding, and testing of expert knowledge. Knowledge engineering was used in LANA-I to design the domain knowledge base, which involve a number of processes: knowledge extraction, knowledge design, and knowledge development.

In the knowledge extraction process, a short interview was conducted with three primary school teachers in Saudi Arabia, who teach science subjects. The aim of this interview was to extract knowledge of lesson design and delivery of tutorials similar to the traditional classroom delivery model. (see Appendix A for the interview questions).

In the knowledge design process, the information from the interviews was then collated and used with the science book to design the tutorials. The designed tutorial was evaluated by the teachers and feedback into the tutorial. Once the tutorials were approved by the teachers, the tutorials were implemented in LANA-I.

In the knowledge implementation process, the domain knowledge base was implemented on a MySQL relational database in order to store and retrieve the resources. The domain knowledge base for the LANA-I consists of two main layers: the domain, which is the science tutorials layer, and a general context layer, illustrated in Figure 4.5.

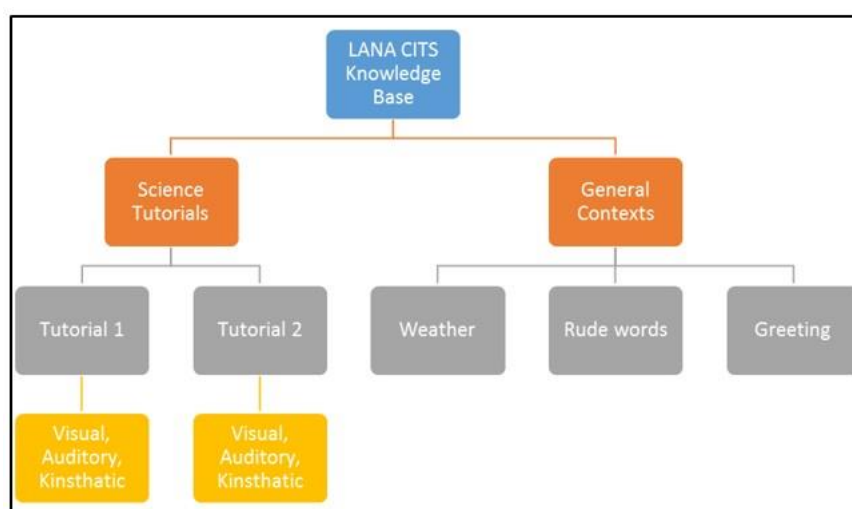


Figure 4.5: The Domain Knowledge base layers

Each layer (Figure 4.5) represents a context, and each context has all the related sub contexts which are mapped using a hierarchical relationship. The domain layer holds all the tutorial sessions related to the science subject, specifically focusing on space subjects such as earth, moon and solar system. The general contexts layer deals with general conversation not related to the domain, such as greeting, weather, rude words, and user leaving words (any words or sentences means that the user will leave the conversation). The general context layer is included into the domain knowledge base to make LANA-I seem more aware by responding to the user's utterances, which are not related to the main domain. However, only a few select sub-contexts have been implemented as a proof of concept within the scope to cover many general chat within this research.

4.4.3 Arabic scripting language

As mentioned in (Chapter 2, section 2.7), there is a lack of Arabic NLP resources leading to limited capabilities within Arabic CAs. Consequently, this research needs a new scripting language to enable the Arabic CA combined with an ITS to deliver Arabic tutoring sessions using the modern standard Arabic language.

As described in (Chapter 2, section 2.7.2), there are three different approaches to develop an Arabic CA and a number of challenges faced by the Arabic language were discussed in (Chapter 2, section 2.7.4). It was concluded that the hybrid similarity approach is considered as a promising approach within the Arabic CA developed in this research, because it combines more than one type of measurements which leads to the similarity between user utterances and patterns being more robust. The foundation of the LANA-I scripting language is based on the InfoChat scripting language (Michie and Sammut, 2001). InfoChat was designed using English scripting language and based on the pattern matching (PM) technique (defined in section 2.2), where the domain is organised into a number of contexts and each context contains rules, each rule in the domain contains a number of patterns and a response that forms the CA output to the user. In the LANA-I CA engine, the similarity strength is calculated through combining the use of the PM, which is based on InfoChat method, and the Cosine algorithm (Qian et al., 2004) to compute the similarity strength between the utterance and the scripted patterns in order to improve the CA accuracy. The highest matching strength pattern will generate the response back to the user.

Based on the scripting methodology reported by (Latham, 2011) the procedures to create the scripts within the Knowledge base are:

1. The methodology for scripting each context is:
 - Create contexts table, which has a record with a unique name to represent that context (topic).
 - An initialisation rule for each context.
 - All rules and patterns are scripted for the associated context.
 - Test each context to check that the expected rule is fired, and check that there is no conflict with the patterns for other rules in the same context.
2. The methodology for scripting each rule is:
 - Create a rule table, which has a unique rule name to represent the tutorial question.

- For each rule, create patterns that match user utterance. Extract the important words and use the wildcards to replace unimportant words.
- Create patterns from each different utterance phrases. e.g. saying the same things using alternative words.
- Create the CA response to the utterance.
- Add Image, Audio, and instructions to the rule to allow mapping to the VAK learning style.

4.4.4 Methodology of Scripting Arabic CA for LANA-I

In utilising the scripting language described in section 4.4.3 for LANA-I, the contexts represent the tutorial topics and the rules represent the agent's questions for such topic, while the pattern represent the user's utterances, which belong to such a rule. The scripting language in LANA-I includes the following novel features:

- Provide supporting material to the user depending on the user's learning style (Visual: images and videos – Auditory: Sounds – Kinaesthetic: Instructions and objects). The learning style will be determined using a child friendly-customised VAK learning style preferences questionnaire before the tutorial.
- This material is stored in the scripting database and once a rule is fired, the corresponding material is provided to the user through the interface. All images, videos and audios provide the right answer. All of them have the same knowledge that the written text provides.
- The scripting language works with the ITS manager to check the user's learning style and provides a suitable material with the fired rule.

Table 4.1 shows an example of the scripting language and how the tutorial is mapped to the VAK model. When the user is visual learner, the rule is fired with the video or image material. In this example (Table 4.1), the rule has video that will be executed along with the rule. When the user is an Auditory learner, the rule will be fired with the audio material. The same thing with the kinaesthetic learning style, the rule will be fired with the instructions. Figure 4.6 shows the models that are used with the Kinaesthetic learning style.

Table 4.1: scripting language

Context: 1 Context Name: Topic 1 (Earth, Sun, and Moon)	Context: 1 Context Name: Topic 1 (الأرض، الشمس، والقمر)
Rule Number: 1 Rule Name: Object moves in the gravitationally curved path called the orbit. The earth rotates around the sun and around its own axis. How long the earth takes to rotate around its axis?	Rule Number: 1 Rule Name: يتكون النظام الشمسي من مجموعة أجرام مثل الأرض والقمر والكواكب. تتحرك الأجرام حول الشمس في مسار يسمى المدار. تدور الأرض حول الشمس وتدور حول محورها فهل تعرف كم تستغرق الأرض لتدور حول محورها؟
Pattern: It * a day Pattern: It * one day Pattern: It * 24 hours Pattern: * twenty four hours	Pattern: * يوم واحد Pattern: * يوم كامل Pattern: * 24 ساعة Pattern: * أربع وعشرون ساعة
Response: Excellent, the earth takes one day (24 hours) to rotate around its axis Image: yes Audio: yes Instruction: Look at the solar system model and its object	Response: أحسنت بارك الله فيك. تستغرق الأرض لتدور حول محورها 24 ساعة أي يوم كامل Image: yes Audio: yes Instruction: أنظر إلى نموذج النظام الشمسي والأجرام المحيطة بها

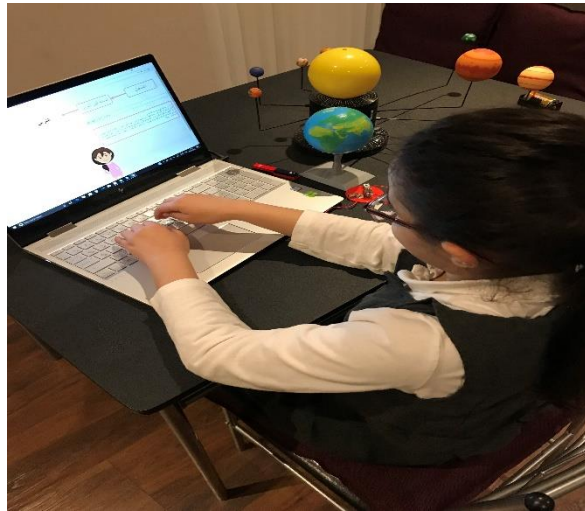


Figure 4.6: the solar system and earth models that are used with the Kinaesthetic learning style

In order to implement LANA-I knowledge base, a database was implemented using MySQL, the database schema is shown on Figure 4.7.

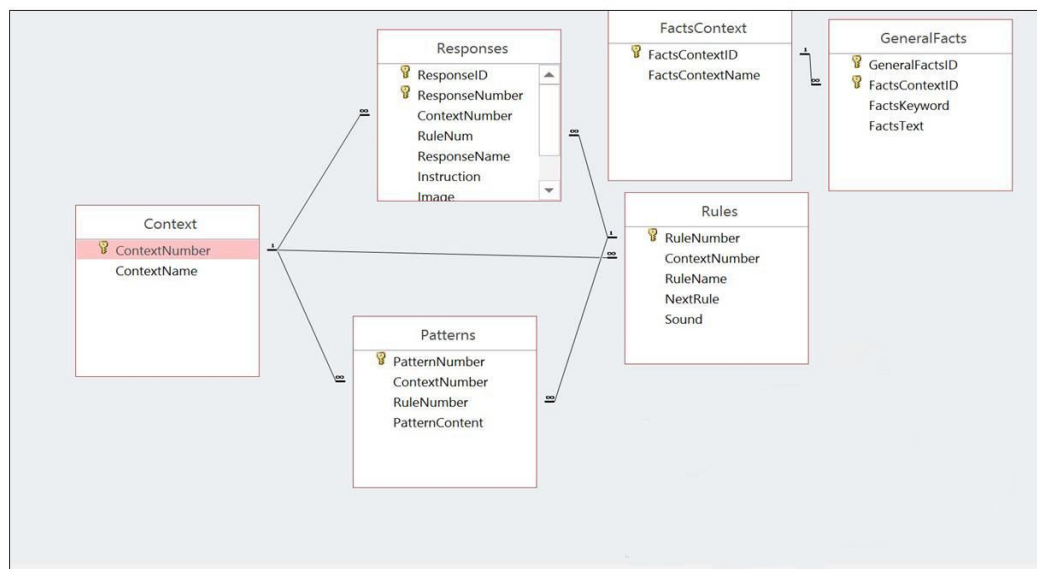


Figure 4.7: Database Schema

4.4.5 VAK learning style questionnaire

In order to personalise the tutorial session according to the student's learning style, a questionnaire will be applied first for each child with a teacher's help. The questionnaire was developed on the basis of a widely disseminated version of (Smith, 1998). This questionnaire focused on Smith's visual, auditory, and kinaesthetic styles (VAK). There were three questions for each style and the pupils had to respond 'yes' or 'no' to each question. The questionnaire was in English and required translation into Arabic. In addition, the wording of the questions had to be modified to be suitable for the reading age of the children used within this study. This translation and modification were done by School Teachers in Saudi Arabia. There is a total of nine questions and the student's learning style result will be based on the highest score in one style. (See the Appendix C for the learning style questionnaire).

4.4.6 Mapping the tutorial into VAK learning style model:

After the tutorial was designed and approved by the teachers then the tutorial questions were mapped to the VAK learning style model. Each tutorial question was associated with its VAK learning style materials. Table 4.2 shows how the tutorial questions for the first topic is mapped to the learning style model VAK.

Table 4.2: Mapping the Tutorial questions to VAK model

Tutorial	VAK
Question: The solar system is the system, which comprising the sun and the objects that orbit it. do you know what objects that exist in the solar system? Answer: Earth, moons and other planets	V: Picture A: Audio K: solar system model
Question: Object moves in the gravitationally curved path called the orbit. The earth rotates around the sun and around its own axis. Do you know the time that the earth takes to rotate around its axis? Answer: A day, or 24 hours.	V: video A: Audio K: earth-moon model
Question: What the result of this rotation? Answer: The day and night	V: Picture A: Audio K: earth-moon model
Question: The sun has many benefits, one of them is to know the time during the day. how we can use the sun to know the time? Answer: The shadow.	V: video A: Audio K: earth-moon model
Question: What cause the four seasons? Answer: Seasons result from the yearly orbit of the Earth around the Sun and the tilt of the Earth's rotational axis relative to the plane of the orbit.	V: video A: Audio K: solar system model
Question: How long the earth takes to complete the rotation around the sun? Answer: 365 days, which is one year	V: video A: Audio K: earth-moon model
Question: The moon is the nearest object to the earth but it is different from the earth in many ways. What are the differences between the earth and the moon? Answer: The moon doesn't have atmosphere and water whereas the earth has.	V: video A: Audio K: earth-moon model
Question: Although the moon can appear a very bright white, its surface is actually dark. So from where the moon reflects the light? Answer: From the sun light.	V: Picture A: Audio K: earth-moon model
Question: We see the moon in different shapes during the month. What is the reason for that? Answer: Because of the moon's rotation around the earth.	V: Picture A: voice K: earth-moon model
Question: What is the duration for the moon to complete one rotation around the earth? Answer: One month, which is 29 Days (based on Arabic calendar)	V: Picture A: Audio K: earth-moon model
Question: The bright part in the moon is changing during the moon's rotation around the earth and this changing is called moon's phases. How many phases does the moon have? Answer: Eight phases.	V: Picture A: Audio K: earth-moon model
Question: What is the astronomical event that occurs to the sun and the moon? Answer: Eclipses	V: Picture A: Audio K: earth-moon model

Each question in the tutorial was mapped to the visual learning style through pictures and videos clips, the auditory learning style through audio clips, and the kinaesthetic

learning style through models and instructions. The pictures and the video clips that are used for the visual learner, were provided from the Saudi Ministry of education website (Education, 2017). All these materials were evaluated and approved by the Primary school teachers in Saudi Arabia. LANA-I will provide the materials based on the student's learning style. For example, if the student is a visual learner, the picture and video clips will be provided to the student during the tutorial.

4.4.7 Log File

LANA-I utilises a temporal memory/log file feature to store several variables in a database table. The variables captured and stored in the log file can be used to evaluate the system. The log file includes two main tables:

1. Conversation log file, which contains:
 - User utterance: the log file captures every utterance that the user typed during the tutorial session.
 - CA generated response: the log file captures all responses generated by the CA.
 - Rules fired during the conversation: all fired rules are captured in the log file. This will help the system to track the conversation and to check if the user reached the tutorial goal or not.
 - Similarity strength: the utterance similarity strength is captured in order to evaluate the similarity algorithm effectiveness.
 - Bad words/utterances: the system will capture any rude words in the log file in order to terminate the session when the user used rude words three times.
 - The time of each utterance: capture the time of each utterance and store it in log file in order to use it to analyse the CA robustness.
2. User's profile: which includes information about the user:
 - The user's first name
 - The gender.
 - The users pre-test and post-test score. A test of ten questions related to the tutorial will be provided to the users before and after the tutorial (see Appendix I).
 - Learning Style.

4.5 Component 2: Designing and Implementing the Arabic CA

The second component illustrated in Figure 4.4 was implemented using the JAVA programming language. The knowledge base and script databases were developed using MySQL. The functions of the individual components of the Arabic CA architecture are described in detail in the following sections.

4.5.1 Controller

The controller is responsible for directing and managing the entire conversation. The controller works with several other components, which are the ITS manager (which will be described in section 4.6.2), and CA manager to achieve the conversational goal. The controller communicates with the ITS manager (component 3, Figure 4.4) to find the student's learning style, which is stored in the database. Then, the controller acquires the user's utterance and provides an utterance checking process. The controller will check the utterance based on the following constraints:

- a) **Cleansing the utterance:** The controller uses the utterance filter to remove special characters (i.e. \$, &, *, !, ?, "", £, (), ^) from the utterance. The filtering ensures that only clean and consistent input is sent forward for pattern matching to make scripting the domain easier.
- b) **Check for rude words:** the system will warn the user three times before terminating the session (O'Shea et al., 2011). The knowledge base database has a table in which all inappropriate words are listed. The list was created through an interview with the primary school teachers. During the interview, a list was created to cover the most unacceptable/inappropriate words that are used by the students in a primary school.

4.5.2 Conversational Agent Manager

The role of the Conversation Manager is to control the conversation to make sure that the goal is achieved, which in this research is the student completing the tutorial. The conversational manager checks whether the user stays on the tutorial topic, or the user switches to another context. Goal-oriented CAs should manage the utterances in an intelligent way (Latham, 2011). For example, when the CA manager receives an utterance that is not related to the tutorial topic, the CA manager checks the user utterance with the general contexts layer to see if the utterance matches other general

contexts within the database. If the utterance matches one of these contexts, the CA manager generates the response to the user, and then the user is brought back to the point where the conversation is interrupted and then directed towards the end of the tutorial.

4.5.3 Conversational Agent Engine

LANA CITS introduces a combination method of STS approach and PM approach to determine the similarity between two sets of strings within CAs, while traditional CAs used only a PM approach that involves strength calculation through different aspects of the user utterance and the scripted pattern such as replacement of selected words by Wildcards. It is important to use the STS approach to overcome some of the challenges in the Arabic language such as the free word order. The primary phase of the engine is based on InfoChat technique (Michie and Sammut, 2001), which used the PM approach to match the user utterance with rules in the stored scripts. The second phase will use the STS algorithm to overcome the PM limitation. In the field of CA development, the scripting is the most time consuming part of CA development (O'Shea et al., 2008). The biggest challenge of scripting CAs is the coverage of all possible user utterances (Latham et al., 2012). The Arabic language is free word order, which means that the Arabic utterance could be written in different word order without affecting the meaning. The main limitation in PM is that it requires the anticipation of all possible user utterances, generation of word order permutations of the utterances and generalization of patterns through the replacement of selected terms by Wildcards. Therefore, it is required in LANA-I engine to handle this challenge by applying the STS algorithm that calculates similarity strength to overcome the scripting challenge. Consequently, strings with word order differences should be recognized as being similar. For example, the sentence: (قرأ المعلم الدرس) and the sentence: (المعلم قرأ الدرس) have the same words and meaning of : (**The teacher reads the lesson**). The difference is the word order, which does not affect the meaning in the Arabic language. When using the PM, both sentences must be scripted in the database in order to be recognised. When using the STS algorithm, it is needed to script one sentence only because the STS algorithm will cover all differences in the word order.

The LANA-I CA engine has two main components that were specifically developed to deal with the features unique to the Arabic language in terms of its morphological nature. These components are:

- Pattern matching approach (Wildcard PM).
- String similarity algorithm (Cosine algorithm)

The combination of these components within LANA-I CA engine will help to calculate the similarity of the user utterance with scripted patterns using string similarity metrics; consequently, reducing the need to cover all possible utterances when scripting the domain.

A. Pattern Matching approach (Wildcard PM)

The first component of the Arabic CA engine, which is the primary phase of the engine, is pattern matching. It is based on similar method to the InfoChat (Michie and Sammut, 2001) of PM, where the user utterance is matched to stored patterns. These patterns contain wild card characters to represent any number of words or characters. For example, the wild card symbol (*) matches any number of words, where the wild card symbol (?) matches a single character. An example for the PM approach in LANA-1 is illustrated in the following table:

Table 4.3: example of matching the utterance with the patterns script

Rule: Although the moon can appear a very bright white, its surface is actually dark. So from where the moon reflects the light?	Rule: القمر لا يضيئ بنفسه فهو انعكاس لضوء آخر. ماهو؟
Pattern 1: the moon reflects * sun Pattern 2: * the sunlight Pattern 3: * the sunshine	Pattern 1: القمر يعكس * الشمس Pattern 2: * ضوء الشمس Pattern 3: * نور الشمس
Utterance 1: it is the sunlight Utterance 2: reflects the moon the sun	Utterance 1: هو ضوء الشمس Utterance 2: يعكس القمر الشمس

In Table 4.3, the utterance#1 will match pattern#2 (* the sunlight), where the wild card symbol (*) will match any number of words. The results from this matching will be 1. Whereas the utterance#2 will not match any patterns because of the word order. In this case the matching result will be 0.

B. Cosine Algorithm

The second component of the Arabic CA engine uses a short text similarity algorithm, which is the Cosine similarity (Qian et al., 2004). The selection of this algorithm was justified in (chapter 2, section 2.7.2.3).

Cosine similarity determines the similarity between two pieces of text by representing each piece of text in the form of word vector. A word vector is a vector of length N where N is the number of different tokens in the text. The similarity is computed as the angle between the word-vectors for the two sentences in vector space.

For two texts t_1 and t_2 the cosine similarity is illustrated in the following equation:

$$SIM(t_1, t_2) = \frac{\sum_{i=1}^n t_{1i} t_{2i}}{\sqrt{\sum t_{1i}^2} \times \sqrt{\sum t_{2i}^2}}$$

Equation 1: Cosine Similarity Equation(Qian et al., 2004)

The cosine similarity result is non negative and bounded between $[0,1]$ where 1 indicates that the two texts are identical.

4.5.4 LANA CA Engine Workflow

First the system will use the PM wildcard to match the user utterance with the pattern saved in the database, if the match is not found, then the next step is to use the STS Cosine similarity. To illustrate the cosine similarity algorithm, assume the user utterance was (t_1), while the pattern stored in the LANA-I database was (t_2). As shown in the following table:

Table 4.4: Cosine Similarity Example

User Utterance: t_1 = القمر ليس له غلاف جوي لذلك يختلف عن الأرض	t_1 : The moon does not have atmosphere therefore it is different from the earth.
Stored pattern in LANA CITS: t_2 = يختلف القمر عن الأرض بأن ليس له غلاف جوي	t_2 : The moon is different from the earth that it does not have an atmosphere.

The utterance in this example (Table 4.4) is not recognised by the PM approach because of the word order or minor differences from the pattern. Therefore, the system applies the Cosine similarity, which is illustrated in the following steps:

- Create Matrix[][] where the columns are the unique words from t_1 and t_2 , and the rows are the words sequence of t_1 and t_2 .

- Table 4.5 shows an example of calculating the similarity of two utterances, where t_1 is the user utterance and t_2 is the stored pattern. The columns are the unique words from t_1 and t_2

- $$\begin{array}{cccccccccc} t_l & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ t_\gamma & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \end{array}$$

- ### Component 3: (LANA-I) Intelligent Tutoring System (ITS) Implementation

LANA-I ITS contains two main components, The Graphical User Interface, and The ITS manager.

4.6.1 Graphical User Interface

The GUI is the point where LANA-I and the user interact with each other. LANA-I has a character called LANA, which appears in all the system interfaces in order to make it more engaging and natural for the users. (Massaro and Bosseler, 2006) developed a computerized tutor for children with autism called (Baladi) to teach vocabulary and functional language use with and without a cartoon face. They evaluated whether the face would increase the rate of learning for receptive measures. The results showed that the children's performance with the face condition was more positive than no face condition. Based on this research, the LANA character was designed and evaluated by primary school teachers in Saudi Arabia in order to make the tutorial more engaging.

The GUI consists of four interfaces: the first interface is the Welcome interface where the user should press the start button to start the session. The second is the pre-test questions interface where the student answers ten questions related to the tutorial. The third is the CA interface, where the GUI takes in user utterances from a textbox and delivers the generated responses back on to the interface. The CA interface is also able to display images, videos, sounds, or instructions depending on the student's learning style. The fourth is the post-test questions interface, which appears after the user complete the tutorial session to evaluate the user learning gain. Figure 4.8 shows screenshot of two interfaces, which are the welcome interface and the tutorial interface (Appendix B shows screenshots of all LANA GUI that have been implemented).



Figure 4.8: LANA-I GUI, Welcome interface and Tutorial interface.

4.6.2 ITS Manager

The ITS manager is responsible for personalising the tutorial according to the user's learning style, which will be determined before the tutorial using a questionnaire (defined in section 4.4.5). The ITS manager can interact with: the conversation agent components through the controller (defined in section 4.5.1), the Knowledge base (defined section 4.4), and the Graphics user interface (section 4.6.1).

4.7 LANA-I Workflow

In this section the system workflow will be described in order to understand how each component communicates with others. Figure 4.9 presents the system workflow via a flowchart. Initially LANA-I will start from the registration screen, the system will ask the teacher to enter their ID (provided by the researcher), age, gender, and the result of the learning style questionnaire, which was conducted before using the system. There are four options for the learning style: Visual, Auditory, Kinaesthetic, and None. The last option will be used with the control group in the testing stage. After completing this stage, the system will show the pre-test interface, and ask the learner

to complete the test. The next interface after submitting the test is the CA tutorial, the ITS manager is responsible in this stage in order to personalise the tutoring session according to the learner style. ITS manager provides the CA components, through the controller, the suitable materials from the Knowledge base component. The ITS manager will also save the learner's registration information and the pre-test score in the log file/student's profile. In the CA screen, the user will be led through the topic from one question to another until the end of the tutoring session. After completing the tutorial, the learner will be asked to fill the post-test questions and the score will be saved by the ITS manager.

Figure 4.10 shows the conversation flowchart, during the conversation, the control will clean the utterance from the special characters (i.e. \$, &, *, !, ?, "", £, (), ^) and will check if the utterance contains a rude word in order to warn the user or terminate session depending on how many times these words are used. The next step in the conversation stage, is to control the conversation to ensure that the goal is achieved and check whether the user stays on the tutorial topic, or whether the user has switched to a general context. The previous step is controlled by the CA manger, which is also responsible for sending the utterance to the CA engine to check the matching. The utterance matching procedure is illustrated in section (4.5.4). The system will generate the response from the Knowledge base component based on the fired rule. The response will be sent to the CA manger, which sends it to the ITS manager through the controller. The ITS manager sent the response to the user through the GUI.

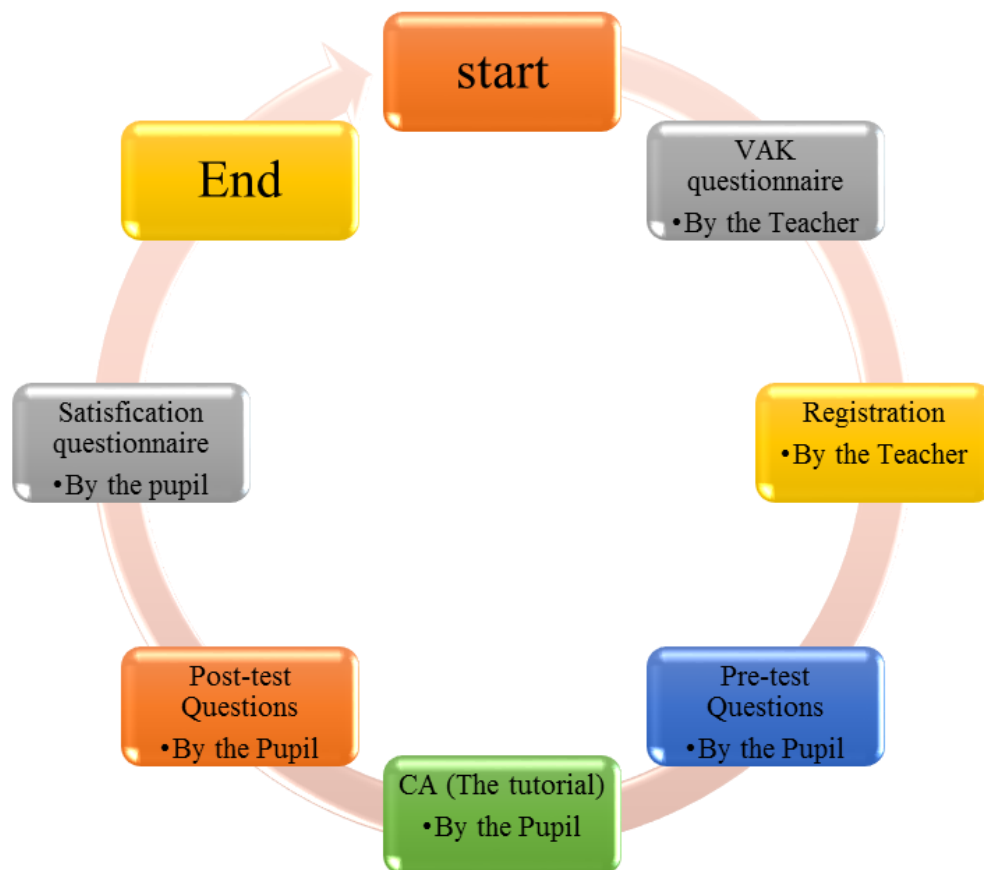


Figure 4.9: LANA-I Workflow

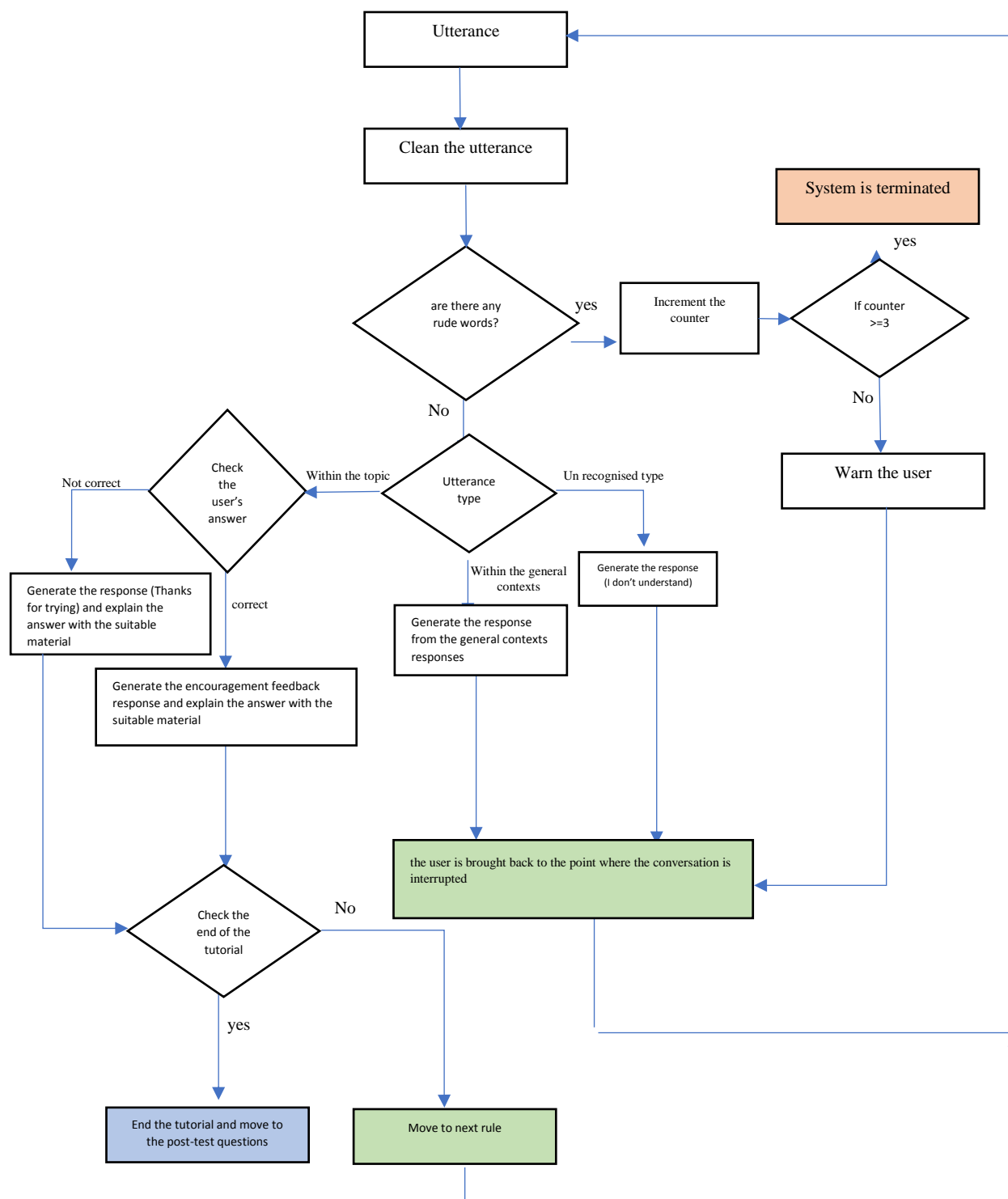


Figure 4.10: LANA-I Conversation Flowchart

4.8 Chapter summary

This chapter has detailed the methodology of implementing the components of LANA-I. These components form the architecture of LANA-I which will be tested for their effectiveness and robustness in order to gather evidence to answer the main research question of ‘Can an Arabic Conversational Intelligent Tutoring System (CITS) be adapted to the learning styles for autistic children and enhance their learning?’ The testing/evaluation methodology, experiments and results are detailed in the next chapter.

The main novel contributions in this chapter are:

- A novel architecture for an Arabic CITS for an appropriate education scenario.
- A novel CA engine that uses the pattern matching technique to find a suitable response to the user’s utterance in the first instance. If no suitable response is found, the engine uses the short text similarity technique to find a response. Combining these techniques reduces the number of scripted patterns and therefore scripting effort.
- Adapting the tutorial to the user’s learning style (VAK) and provide suitable material to the user according to the user’s learning style (images and videos for Visual learner, Sound for Auditory learner, physical models for Kinaesthetic learner).

Chapter 5 – LANA-I Evaluation Methodology and Results

5.1 Overview

In this chapter the evaluation methodology of the first LANA CITS (LANA-I) is proposed in order to evaluate the effectiveness of the LANA-I framework and architecture. In Chapter 4 a novel LANA-I architecture was proposed and utilised to implement LANA-I, which was designed to deliver topics related to a science subject by engaging with the user (age 10-12 years old) in the Arabic language. LANA-I uses the Visual, Auditory, and Kinaesthetic (VAK) learning style model to personalise the lesson according to the user learning style preference. The architecture of LANA-I incorporates several components such as the similarity algorithm, mapping the tutorial to the VAK learning style model, and the Arabic scripting language.

The following sections outline the evaluation methodology through two experiments. The aim of the first experiment is to answer two questions related to Hypothesis A (LANA-I using VAK model can be adapted to the student learning style), which are:

Question1: Can LANA-I which embeds the learning Style VAK help a student to improve their learning?

Question2: Can LANA-I deliver personalised adaptation of a tutorial to the student's learning style successfully?

The second experiment was conducted to test Hypothesis B (LANA-I is an effective Arabic CA) designed to answer the following questions:

Question 1: Are the students satisfied with the usability of LANA-I?

Question 2: Is the Arabic conversational agent used within LANA-I robust?

In order to measure these hypotheses, a set of objective and subjective metrics were defined and justified. The objective metrics were measured through the log file/temporal memory (Section 5.9). The subjective metrics were tested using the user feedback questionnaire (Section 5.2.2). The aim of capturing the subjective and objective metrics is to test both hypotheses, which are related to: the effectiveness of the conversation and using the VAK learning style, end user satisfaction, usability and system robustness.

5.2 (LANA-I) Evaluation Methodology

The effectiveness of LANA-I was tested through a pilot study conducted with neurotypical participants in the target age group (10-12) years old whose first language is Arabic. None of the children had been diagnosed as Autistic. It is important to test the quality of the tutorial and the materials, which are used in LANA-I, with the general population before testing it with Autistic children. The system robustness and its ability to complete the task must be tested with general population before autistic children in order to avoid any confusion may be happened to the autistic child during the tutorial. The results of the pilot study will then be used to adjust the full-scale experiment to ensure its accuracy. The tutorial used in this evaluation was the (solar system tutorial), which was designed in Chapter 4, section 4.4.6.

5.2.1 Hypothesis

The main hypotheses of the experiments were:

- **HA₁:** LANA-I using VAK model can be adapted to the student learning style:

This hypothesis relates to the VAK adaptation and the system ability to adapt to the student learning style.

HA₀: LANA-I using VAK model cannot be adapted to the student learning style:

The conversational intelligent tutoring system and the learning style VAK, which have been selected are not suitable to adapt the student's learning style.

- **HB₁:** LANA-I is an effective Arabic CA

This hypothesis relates to the system robustness and its ability to complete the task. The LANA-I CA engine, the Arabic scripting language, and methodologies deployed result in an effective functional Arabic CA. (see section 5.3 for metrics to measure the effectiveness).

- **HB₀:** LANA-I is not an effective Arabic CA

The Arabic CA engine, the Arabic scripting language, and methodologies deployed result cannot provide an effective Arabic CA.

Both HA and HB were tested by the subjective and objective user evaluation.








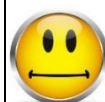





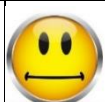





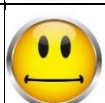



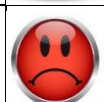
5.2.2 Subjective user evaluation

The subjective evaluation was measured using a user evaluation feedback questionnaire, which was designed to identify the qualitative user feedback after the LANA-I tutorial.

Table 5.1 shows the feedback questionnaire, which contains eight questions to be rated using the three face scale (happy, natural, sad) (Reynolds-Keefer and Johnson, 2011). It is reported that questionnaires with three face scale are more understandable in some situation where the ability of reading is difficult such as with younger participants (Zhang et al., 2002). Many children cannot give a verbal answer of their opinion or feeling, either because they are young, or because of neurological disorder or communication challenging such as Autistic children (Yahaya and Salam, 2016). In this situation the using of three face scale will be useful and easy to use especially with Autistic children.

The questionnaire was designed to suite the target age group and also make the scale clearer and more comprehensible. The questions were designed in Arabic for children in the targeted age range and were evaluated and approved by three primary school teachers in Saudi arabia. The questions were also designed to extract the user's opinion on the performance and usability of LANA-I by selecting the most appropriate face representing their feeling against each of the eight questions. The participants completed the feedback questionnaire, in Arabic, after completing the tutorial session. The teacher and the researcher were present whilst they completed the questionnaire to answer any questions.

Table 5.1: User Feedback Questionnaire (Translated version)

1. I think the LANA Tutor is enjoyable to use			
2. I think LANA had the right amount of information in one Tutorial.			
3. I think I remember what LANA has taught me in the tutorial.			
4. I think I can learn better when using my learning style (Visual – Auditory -Kinaesthetic).			
5. I like the design and colour of LANA.			
6. I think I can learn from LANA better than my teacher.			
7. I think I would like to use LANA again with other lessons.			
8. I think talking with LANA is like talking with my friend.			

5.2.3 Objective user evaluation

The objective evaluation was achieved through analysing the LANA-I log file, which stores interaction information from the user during the tutoring session. The log file stores a number of variables such as the rule fired, the timing of the tutorial, the utterance, and the response. This data was analysed to explore the effectiveness of LANA-I CA (Section 5.9).

5.3 Design of evaluation metrics

With the purpose of measuring user feedback and evaluation presented in the previous section, a mechanism called a goal question metric approach (GQM) was used (Fenton and Pfleeger, 1998) (Kaleem et al., 2016). This mechanism is based on the supposition that goals must be identified for a system then number of questions is defined for each goal, and a set of metrics associated with each question (Fenton and Pfleeger, 1998).

The (GQM) model was used to identify which metrics were required to test the hypothesis successfully. Figure 5.1 shows that there are two goals related to

hypotheses and each goal has two questions. These questions used a combination of subjective metrics (questionnaire metrics), and objective metrics (log file metrics).

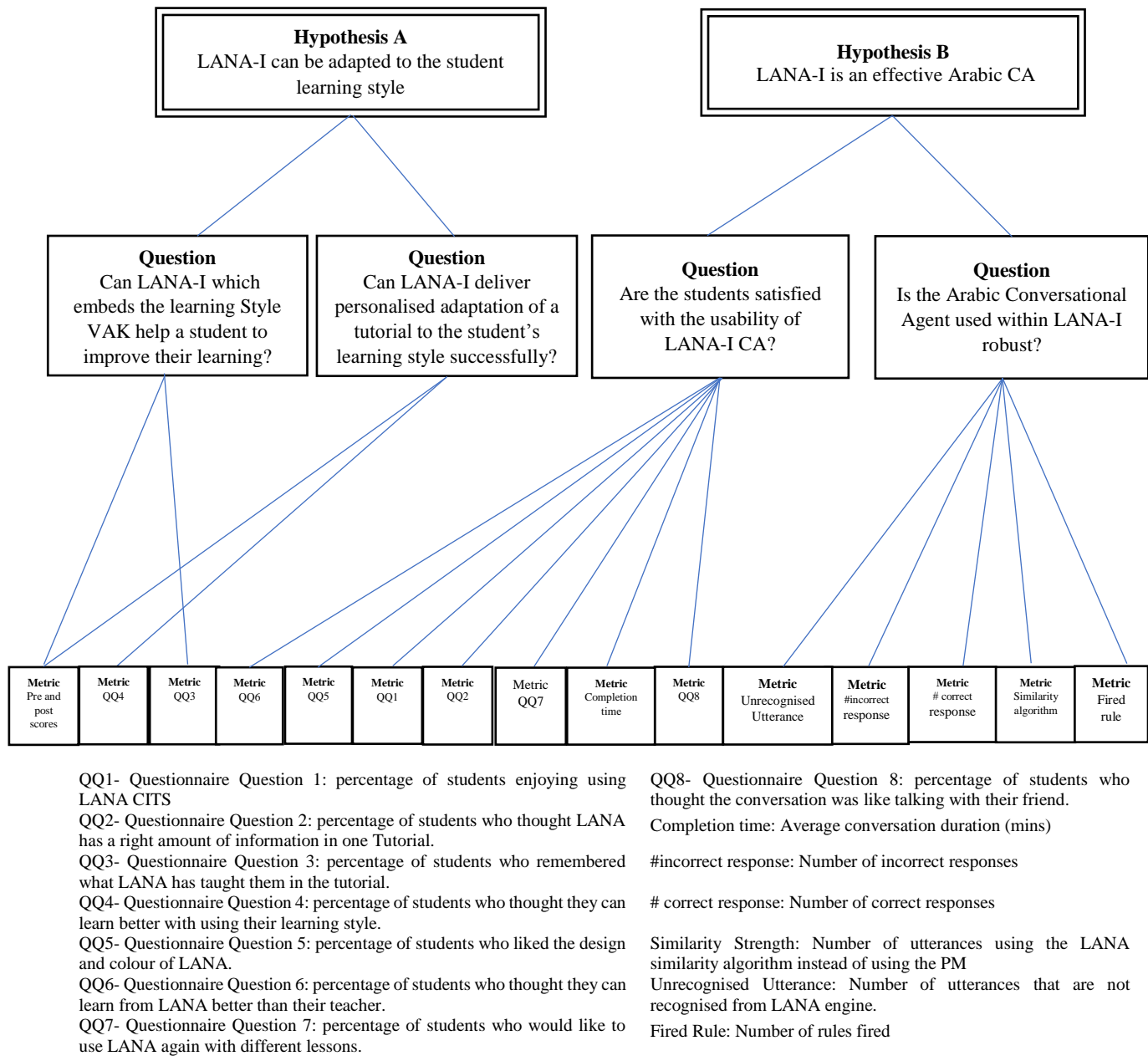


Figure 5.1: GQM Model for LANA-I

5.3.1 Experimental Methods for Subjective and Objective Evaluation

Subjective metrics were measured by gathering the qualitative user feedback. The participants were asked to complete a short user evaluation questionnaire (see section 5.2.2), after completing the tutorial session. The subjective metrics, and the methods used with each metric and the related hypothesis are illustrated in Table 5.2. The

objective metrics (see section 5.2.3) were measured through analysis of the LANA-I log file. Table 5.3 shows the objective metrics and their evaluation methods and the related hypothesis.

Table 5.2: Subjective Evaluation Metrics

Subjective Metrics		
Metrics to be Evaluated	Method of Evaluation	Hypothesis
QQ1: LANA Tutor is enjoyable to use	Questionnaire	Hypothesis B
QQ2: LANA has a right amount of information in one Tutorial.	Questionnaire	Hypothesis B
QQ3: I remember what LANA has taught me in the tutorial.	Questionnaire	Hypothesis A
QQ4: I can learn better with using my learning style.	Questionnaire	Hypothesis A
QQ5: I like the design and colour of the interface.	Questionnaire	Hypothesis B
QQ6: I can learn from Lana better than learning from the teacher.	Questionnaire	Hypothesis B
QQ7: I would like to use LANA again with different topics.	Questionnaire	Hypothesis B
QQ8: The conversation was natural.	Questionnaire	Hypothesis B

Table 5.3: Objective Evaluation Metrics

Objective Metrics		
Metrics to be Evaluated	Method of Evaluation	Hypothesis
Tutorial Completion time	Log file	Hypothesis B
Number of unrecognised Utterances	Log file	Hypothesis B
Number of correct responses	Log file	Hypothesis B
Number of incorrect responses	Log file	Hypothesis B
Number of utterances using Similarity algorithm	Log file	Hypothesis B
Number of rules fired during tutorial	Log file	Hypothesis B
Pre-test and post-test scores	Log file	Hypothesis A

5.4. Participants and Ethical Consideration

The total sample size is 24 neurotypical participants aged between 10-12 years old and their first language is Arabic. The participants recruited for the evaluation are residents of the Greater Manchester Area, UK and none of them had any previous experience using LANA-I. All participants' parents received an information sheet about the project and its aims, and a consent form to get their permission before conducting the experiment (see Appendix D and E for the information sheet and consent form). The participants were divided into two equal sized groups. The first group was the control group (12 participants) who used LANA-I without adapting to the learning style VAK model. The second group was the experimental group (12 participants) used LANA-I

with the adapting to the learning style VAK model. The reason of dividing the participants into two groups was to test the Hypothesis A (LANA-I can be adapted to the student learning style).

5.5. Participants Interaction

5.5.1 Experiment 1: LANA-I Tutoring with and without adapting to VAK learning Style

The aim of this experiment is to test the hypothesis HA_1 based on the metrics outlined in section 5.3. The experiment will compare the pre-test and post-test scores of participants with and without adapting to the VAK learning Style model. The subjective and objective metrics (Figure 5.1 GQM Model for LANA-I), were used to answer the two questions related to Hypothesis A (LANA-I can be adapted to the student learning style), which are:

Question1: Can LANA-I which embeds the learning Style VAK help a student to improve their learning?

Question2: Can LANA-I deliver personalised adaptation of a tutorial to the student's learning style successfully?

Each group of the participants were asked first to register into the system and complete the pre-test questions. The control group used LANA-I without adapting to the VAK learning style model. They started the tutorial without VAK questionnaire, whereas the experimental group, who used the LANA-I with adapting to VAK learning style model, were asked to fill the VAK learning style questionnaire in order to adapt their learning style. After adapting to the learning style, the tutorial provided the suitable material during the session such as videos, images or instructions. The control group used the tutorial session based on a text conversation without any materials. When the session ended, both groups did the post-test questions in order to measure their learning gain (chapter 4, section 4.4.7). Finally, both groups had to complete the user feedback questionnaire.

5.5.2 Experiment 2 – LANA-I CA System Robustness

This experiment was conducted to test the hypothesis HB_1 based on the metrics outlined in section 5.3. The data for this experiment was gathered from the LANA-I log file and the user feedback questionnaire. The subjective and objective metrics

(Figure 5.1 GQM Model for LANA-I), were used to answer questions related to Hypothesis B (LANA-I is an effective Arabic CA):

Question 1: Are the students satisfied with the usability of LANA-I?

Question 2: Is the Arabic Conversational Agent used within LANA-I robust?

The data gathered from the log file allowed assessment of the performance of LANA-I CA and the algorithms deployed in the architecture. This data measured the success of the objective metrics (see Table 5.3 for list of objective metrics). The data from user feedback questionnaire was analysed in order to measure the success of the subjective metrics (see Table 5.2 for list of subjective metrics).

5.6 Experimental Data Analysis

The data gathered from the participant interactions in experiment 1 and 2 was collated, and analysed in order to explore the findings. The data was analysed using the software SPSS and Microsoft Excel. The analysed data was used to highlight weaknesses and areas for improvement within the system.

5.7 Data Normality Testing

A normality test was used to determine whether the data has been drawn from a normally distributed population (Moore et al., 2012). It is important to test the normality of the data to choose the right test. If the data is normal distribution, a parametric test will be used. If the data is not normal distribution, a non-parametric test will be used. Tests of normality can be done by using Kolmogorov-Smirnov test for a sample size greater than 50 or Shapiro-Wilk test if sample size is smaller than 50 (Moore et al., 2012). In order to know whether the data is normal distribution or not, the significant values greater than 0.05 indicates that sample scores are similar to a normal distribution. Appendix F shows the histograms for the questionnaire results and the pre-test and post-test scores along with the results of the normality test. The histograms and the Shapiro-Wilk test shows that the data is not normally distributed. Therefore, non-parametric tests will be used to analyse the data.

5.7.1 Wilcoxon Signed Ranks Test

The Wilcoxon Signed Ranks Test (Moore et al., 2012) is an alternative to the parametric test (paired dependent t-test) - to compare two sets of scores that come

from the same participant. In the first experiment, the pre-test and post-test scores were taken from the same participant. It assesses whether the mean ranks for the two scores differ significantly. The two outputs that must be interested in are the Z value and the associated significance levels (must be less than P-Value = 0.05), presented as Asymp. Sig. (2-tailed). If there is a statistically significant difference between the groups, the direction of the difference must be presented by describing which group is higher from the Ranks table under the column Mean Rank. However, it is very important to provide an effect size statistic with the results even if the P-Value was significant. Effect size statistics is defined as *“information about the magnitude and direction of the difference between two groups or the relationship between two variables”* (Durlak, 2009). The Effect size will be measure according to (Cohen, 1988) criteria of 0.1 = small effect, 0.3 = medium effect, 0.5 = large effect. In Wilcoxon test, the value of Z can be used to calculate an approximate value of effect size (r).

$$r = Z / \text{square root of } N \text{ where } N = \text{total number of cases.}$$

Equation 2: effect size calculation (Moore et al., 2012)

5.7.2 Mann-Whitney U test

Mann-Whitney U is used to test for differences between two independent groups on a continuous measure (Moore et al., 2012). This technique is used as an alternative to the t-test for independent samples. It is a non-parametric test to compare the medians of the two groups instead of comparing the means of the two groups, as in the case of the t-test. Mann-Whitney U test converts the scores on the continuous variable to ranks, across the two groups to evaluate whether the ranks for the two groups differ significantly. As same as Wilcoxon test, the two main outputs that must be interested in are the Z value (to calculate the effect size) and the associated significance levels (P-Value must be less than 0.05), presented as Asymp. Sig. (2-tailed). The effect size is calculated using the same equation (Equation 2).

5.7.3 Chi-Square Test of Independence

This test is used to explore the relationship between two categorical variables. Each of these variables can have two or more categories. In this test, the observed frequencies of cases that occur in each of the categories, are compared with the values that would be expected. Three tables will be created from this test, (1) a cross

tabulation table, which shows the cases classified according to the categories in each variable. The cross tabulation table illustrate the percentage between the categories in each variable. (2) Chi-square table, which has the main value Asymp. Sig. (2-sided), which shows the P-Value (the significant level). (3) Symmetric Measures table, which used to find out the effect size (r) that is available in this table in the variable **phi coefficient**. In this experiment, the Chi- Square test was used in some of the questionnaire's questions to explore if there is a relationship between the user's perception and the VAK learning style.

5.8 Experiment 1: results and discussion

The aim of this experiment is to answer two questions related to Hypothesis A, which are:

Question1: Can LANA-I which embeds the learning Style VAK help a student to improve their learning?

Question2: Can LANA-I deliver personalised adaptation of a tutorial to the student's learning style successfully?

The experiment is based on the subjective and objective metrics outlined in section 5.3. The data was gathered from the user feedback questionnaire and the pre-test and post-test scores with and without adapting to the VAK learning Style.

The learning gain was measured using a pre-test and post-test approach (Kelly and Tangney, 2006; Graesser et al., 2005; Lee et al., 2006) (chapter 2, section 2.8.3). The same test questions were completed before and after the tutoring conversation. The test scores were compared to establish whether there is any improvement as follows:

$$\text{Learning gain} = \text{post-test score} - \text{pre-test score}.$$

Equation 3: Learning gain

Tables 5.4 and 5.5 show the pre-test, post-test, and the learning gain with and without adapting to the VAK learning style.

Table 5.4: Pre-test and Post-test scores in Control Group

Participant	Control Group		
	Pre-test (out of 10)	Post-test (out of 10)	Learning gain
1	6	8	2
2	6	9	3
3	4	8	4
4	6	8	2
5	5	7	2
6	4	8	4
7	4	6	2
8	7	8	1
9	9	9	0
10	4	6	2
11	4	8	4
12	6	9	3
The Mean	5.4	7.8	2.4

Table 5.5: Pre-test and Post-test scores in Experimental Group

Participants	Learning Style	Experimental Group		
		Pre-test (out of 10)	Post test (out of 10)	Learning gain
1	A	5	7	2
2	A	5	8	3
3	A	5	7	2
4	A	4	7	3
5	K	6	10	4
6	K	6	10	4
7	K	5	10	5
8	K	6	10	4
9	V	6	10	4
10	V	6	9	3
11	V	4	8	4
12	V	6	10	4
The Mean		5.3	8.8	3.5

Table 5.6 and Table 5.7 show the results of Wilcoxon Signed Ranks test carried out to determine if there was any difference between the pre-test and post-test scores in the Control Group.

Table 5.6: The difference between the mean of the pre-test and post-test scores in the Control Group

Descriptive Statistics								
	N	Mean	Std. Deviation	Minimu m	Maximu m	Percentiles		
						25th	50th (Median)	75th
post_test	12	7.8333	1.02986	6.00	9.00	7.2500	8.0000	8.7500
pre_test	12	5.4167	1.56428	4.00	9.00	4.0000	5.5000	6.0000

Table 5.7: The P Value between the means from Table 5.6

Test Statistics ^a	
	pre_test - post_test
Z	-2.971 ^b
Asymp. Sig. (2-tailed)	.003

a. Wilcoxon Signed Ranks Test

b. Based on positive ranks.

When comparing the mean ranks in Table 5.6 of the two scores, the post-test scores was ranked higher, which means that the students performed well in the post-test in the control group. As Table 5.7 shows p-values less than 0.01 ($p = .003$), which implies that the students performed significantly better in the post-test. The effect size (r) is calculated using the (Equation 2) and found that $r = 0.6$ indicating a large effect size using (Cohen, 1988) criteria.

Table 5.8 and Table 5.9 show the results of Wilcoxon Signed Ranks test carried out to determine if there was any difference between the pre-test and post-test scores in the Experimental Group.

Table 5.8: The difference between the mean of the pre-test and post-test scores in the Experimental Group

Descriptive Statistics								
	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
post_testVAK	12	8.8333	1.33712	7.00	10.00	7.2500	9.5000	10.0000
pre_testVAK	12	5.3333	.77850	4.00	6.00	5.0000	5.5000	6.0000

Table 5.9: The P Value between the means from Table 5.8

Test Statistics ^a	
	pre_testVAK - post_testVAK
Z	-3.108 ^b
Asymp. Sig. (2-tailed)	.002

a. Wilcoxon Signed Ranks Test

b. Based on positive ranks.

The results from Table 5.8 and Table 5.9 show that there is a statistically significant difference between the pre-test and post-test scores with p-values less than 0.01 ($p = .002$) and the mean value in the post-test is greater than the mean in the pre-test. The effect size (r) is calculated using the (Equation 2) and found that $r = 0.6$ indicating a large effect size using (Cohen, 1988) criteria. It can be concluded that there is a statistically significant difference between the pre-test and post-test scores in the Experimental Group.

Table 5.10 illustrates the results of the Mann-Whitney U test conducted in order to measure whether there was a statistically significant difference between the learning gain between the two groups (Control Group: without using VAK, Experimental Group: with using VAK).

Table 5.10: Mann-Whitney U test for the learning gain between Control Group and Experimental Group

	Group	N	Mean Rank	Sum of Ranks
Learning gain	Control	12	9.50	114.00
	Experimental	12	15.50	186.00
	Total	24		
Test Statistics^a				
		Learning gain		
Mann-Whitney U				36.000
Wilcoxon W				114.000
Z				-2.173
Asymp. Sig. (2-tailed)				.030
Exact Sig. [2*(1-tailed Sig.)]				.039 ^b
a. Grouping Variable: VAK				
b. Not corrected for ties.				

Table 5.10 shows that there is a difference in the mean value between the Control Group learning gain and Experimental Group learning gain. The results show that the Experimental Group learning gain were improved more than the Control Group. The learning gain mean ranks in each case are Control Group ($M = 9.50$) and Experimental Group ($M = 15.50$). The learning gain between Control Group and Experimental Group is statistically significant different as the p-value is less than 0.05 ($P = .03$). The effect size (r) is calculated using the (Equation 2) and found that $r = 0.4$ indicating a medium to large effect size using (Cohen, 1988) criteria. It can be concluded that there is a statistically significant improvement in the learning gain in the Experimental Group comparing to the Control Group.

Relative learning gain was also measured in this experiment. Relative learning gain is a measure that calculate the average improvement in test scores as a percentage of the possible improvement (Latham et al., 2014). This measure additionally takes into account the opportunity for improvement. Average test score improvements were calculated and compared using the following formula:

$$\text{Relative learning gain} = (\text{PostTest} - \text{PreTest}) / (\text{TotalPossibleScore} - \text{PreTest})$$

Equation 4: Relative Learning Gain (Latham et al., 2014)

For example, if a student gets 8/10 in pre-test and only improves by 1, this is different to another getting 3/10 in pre-test and only improving by 1 – improvement is 50% in first case but only 14.2% in the second case.

Table 5.11 illustrates the results of the Mann-Whitney U test conducted in order to measure the relative learning gain between Control Group and Experimental Group.

Table 5.11: Mann-Whitney U test for the relative learning gain between Control Group and Experimental Group.

	Group	N	Mean Rank	Sum of Ranks
Relative Learning gain	Control	12	9.04	108.50
	Experimental	12	15.96	191.50
	Total	24		
Test Statistics^a				
		Relative Learning gain		
Mann-Whitney U				30.500
Wilcoxon W				108.500
Z				-2.428
Asymp. Sig. (2-tailed)				.015
Exact Sig. [2*(1-tailed Sig.)]				.014 ^b
a. Grouping Variable: VAK				
b. Not corrected for ties.				

Table 5.11 shows that there is a difference in the mean value between Control Group and Experimental Group. The average improvement in test scores in the Experimental Group were improved more than the Control Group. The ranks in each case are: Control Group (M= 9.04) and Experimental Group (M= 15.96). The relative learning gain between the Control Group and the Experimental Group is statistically significant different, p-value less than 0.05 (P= .015). The effect size (r) is calculated using the (Equation 2) and found that $r = 0.5$ indicating a large effect size using (Cohen, 1988) criteria. It can be concluded that the average improvement in test scores in the Experimental Group is statistically significant better than the Control Group.

This result highlight that **the participants performed better when they are adapting to their preferred VAK learning style**. It can therefore be concluded that there is statistically a significant difference between the pre-test and post-test scores with and without adapting to VAK learning style, indicating that **HA₁ can be accepted**.

Another test was conducted to find out whether there was a significant difference in the scores for participant's **perception of remembering what they have learned from LANA-I**. This was conducted between the Control Group (learning without adapting to VAK learning style) and the Experimental Group (learning with adapting to VAK learning style). The data was gathered from the user feedback questionnaire (Question 3: I think I remember what LANA has taught me in the tutorial). The results illustrated in Table 5.12 indicated that participants who learned by adapting to their VAK learning style (Experimental Group) were much happier with remembering what they have learned from LANA-I (91.7%) than participants who learned without adaptation to VAK learning style (50%) (Control Group). This result indicates that **HA₁ can be accepted**.

Table 5.12: Crosstab * VAK - Question 3 (I think I remember what LANA has taught me in the tutorial)

			VAK		Total
			With	Without	
I think I remember what LANA has taught me in the tutorial	Happy	Count	11	6	17
		% within VAK	91.7%	50.0%	70.8%
	Normal	Count	1	6	7
		% within VAK	8.3%	50.0%	29.2%
Total		Count	12	12	24
		% within VAK	100.0%	100.0%	100.0%

Table 5.13 shows the Chi-Square Tests for (Question 3: I think I remember what LANA has taught me in the tutorial) with and without adapting to VAK learning style. There is a statistically significant difference in answering this question, $p = .025$ (less than .05), Table 5.14 shows the effect size in this test, $\Phi = 0.45 \approx 0.5$ indicating a medium to large effect size using (Cohen, 1988) criteria, meaning that **HA₁ can be accepted**.

Table 5.13: Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	5.042 ^a	1	.025		
Continuity Correction ^b	3.227	1	.072		
Likelihood Ratio	5.455	1	.020		
Fisher's Exact Test				.069	.034
N of Valid Cases	24				

a. 2 cells (50.0%) have expected count less than 5. The minimum expected count is 3.50.

Table 5.14: Testing the effect size

Symmetric Measures			
		Value	Approx. Sig.
Nominal by Nominal	Phi	.458	.025
	Cramer's V	.458	.025
N of Valid Cases		24	

Further tests have been carried out to find out whether there was a significant difference in the scores for participant's **satisfaction with adapting the tutorial to the VAK learning style**. The data was gathered from the user feedback questionnaire (Question 4: I think I can learn better with adapting to my learning style (Visual – Auditory -Kinaesthetic). The results illustrated in Table 5.15 that participants are happy with (83.3%), indicating that **HA₁ can be accepted**.

Table 5.15: Crosstab * using VAK - Question 4 (I think I can learn better with adapting to my learning style)

			VAK	Total
			With	
I think I can learn better with using my learning style (Visual – Auditory -Kinaesthetic)	Happy	Count	10	10
		% within VAK	83.3%	83.3%
	Normal	Count	2	2
		% within VAK	16.7%	16.7%
Total		Count	12	12
		% within VAK	100.0%	100.0%

Table 5.16 shows the results of Wilcoxon Signed Ranks test conducted to test the students' performance between the three learning styles (V: Visual learning style, A: Auditory learning style, K: kinaesthetic learning style).

Table 5.16: The differences between the mean ranks between the three learning styles (V: visual, A: auditory, K: kinaesthetic)

Descriptive Statistics								
	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
pre_testA	4	4.7500	.50000	4.00	5.00	4.2500	5.0000	5.0000
pre_testV	4	5.5000	1.00000	4.00	6.00	4.5000	6.0000	6.0000
pre_testK	4	5.7500	.50000	5.00	6.00	5.2500	6.0000	6.0000
post_testA	4	7.2500	.50000	7.00	8.00	7.0000	7.0000	7.7500
post_testV	4	9.2500	.95743	8.00	10.00	8.2500	9.5000	10.0000
post_testK	4	10.0000	.00000	10.00	10.00	10.0000	10.0000	10.0000

Each of the three learning styles has differences between the means (the pre-test and post-test scores). The mean of the post-test scores are higher for all the learning style (V, A, and K) than the pre-test scores which means that the students performed well in the post test after adapting to their learning style. However, Table 5.17 shows that there is not a statistically significant difference between the pre-test and post-test scores for each learning style and this is because of the small sample size.

Table 5.17: The P-Values between the means from table 5.16.

Test Statistics ^a			
	post_testA - pre_testA	post_testV - pre_testV	post_testK - pre_testK
Z	-1.857 ^b	-1.890 ^b	-1.890 ^b
Asymp. Sig. (2-tailed)	.063	.059	.059
a. Wilcoxon Signed Ranks Test			

5.9 Experiment 2: results and discussion

This experiment was conducted to answer two questions related to Hypothesis B, which are:

Question 1: Are the participants satisfied with the usability of LANA-I?

Question 2: Is the Arabic Conversational Agent used within LANA-I robust?

The experiment is based on the subjective and objectives matrices outlined in section 5.3. The data was gathered from the LANA-I log file and the user feedback questionnaire. An example translated transcript is illustrated in Table 5.18. The participant in the example was given the tutorial that is adapted to Kinaesthetic learning style.

Table 5.18: Excerpt from LANA Log file transcript

		English translation	Arabic Utterance
1	LANA	Welcome to LANA, the tutoring system. The solar system is the system, which comprising the sun and the objects that orbit it. Look at solar system model. Tell me what objects that exist in the solar system?	مرحبا بك في برنامج لانا التعليمي النظام الشمسي هو النظام الذي يضم الشمس ومجموعة من الاجرام التي تدور حولها. انظر الى نموذج النظام الشمسي امامك وأذكر لي مم يتكون؟
2	User	The stars and the planets	النجوم والكواكب
3	LANA	Well done, look at the solar system model, the solar system contains many objects such as the earth, moon and planets.	احسنت بآرك الله فيك انظر الى نموذج النظام الشمسي امامك يتكون النظام الشمسي من مجموعة أجرام مثل الأرض والقمر والكواكب تتحرك الأجرام حول الشمس في مسار يسمى بالمدار. تدور الأرض حول الشمس وتدور حول

		Object moves in the gravitationally curved path called the orbit. The earth rotates around the sun and around its own axis. How long does it take the earth to rotate around its axis?	محورها فكم تستغرق الأرض لتدور حول محورها؟
4	User	24 Hours	24 ساعة
5	LANA	<i>Well done</i> , the earth takes 24 hours to rotate around its axis. What is the result from this rotation?	<i>احسنت بارك الله فيك</i> تستغرق الأرض لتدور حول محورها 24 ساعة أي يوم كامل ماذا ينتج عن هذا الدوران؟
6	User	Day and night	الليل والنهار
7	LANA	<i>Well done</i> , Using the earth model and the light, turn on the light and face it to the earth model (the light here is as the model of the sun). Now you will see that the half of the earth is shine and the half is dark. The sun has many benefits; one of them is to know the time during the day. how we can use the sun to know the time?	<i>احسنت بارك الله فيك</i> استخدام نموذج الكرة الأرضية والمصباح. أولا أشعل المصباح ووجهه إلى نموذج الكرة الأرضية (المصباح هنا يمثل الشمس) والآن ستري أن نصف الكرة الأرضية مضاء ونصفها الآخر معتم. وهذا يمثل الليل والنهار الشمس لها منافع كثيرة منها تقدير الوقت خلال النهار فكيف نستطيع ذلك؟
8	User	The Shadow	الظلال
9	LANA	<i>Well done</i> , Use the light and a pen to create a shadow. A shadow is a space where light from a light source is blocked by an opaque object. Now we will learn how to use shadow to know the time during the daytime: the length of the shadow is changed depends on to the changing of the sun. The shadow's length is long on the morning, shortest on the afternoon, then the length is increased in the other side in until the sunset. The earth rotates around the sun. what the result of this rotation?	<i>احسنت بارك الله فيك</i> . استخدم المصباح والقلم أمامك لإظهار ظل القلم. نتعرف على الوقت من خلال الظل ويكون الظل عندما يعترض جسم ما مسار اشعة ضوئية فلا يستطيع الضوء المرور عبر ذلك الجسم ويكون نتيجة لذلك جهة معتمة اسمها الظل. ويتغير طول الظل حسب اتجاه الشمس فيكون طول الظل كبير في النهار ويقصر في الظهيرة ويكون أقصر مايمكن عند الغروب. تدور الأرض حول الشمس كل سنة. ماهي نتيجة هذا الدوران؟
10	User	Four seasons	الأربع مواسم
11	LANA	<i>Well done</i> , Use the solar system model and orbit the earth around the sun. the earth rotates around the sun every year and this cause the four seasons. How long the earth takes to complete the rotation around the sun?	<i>احسنت بارك الله فيك</i> • استخدم نموذج النظام الشمسي وأجعل الأرض تدور حول الشمس تدور الأرض حول الشمس كل سنة وينتج عن هذا الدوران الفصول الأربعة كم يستغرق دوران الأرض حول الشمس؟
12	User	365 days	365 يوم
13	LANA	<i>Well done</i> , the earth takes 365 days to complete one rotation around the sun. <i>Look at the earth model, you will see a small object, this is the moon.</i> The moon is the nearest object to the earth but it is different from the earth in many ways. What are the differences between the earth and the moon?	<i>احسنت بارك الله فيك</i> يستغرق دوران الأرض حول الشمس 365 يوما انظر إلى نموذج الكرة الأرضية ستري جرم صغير بجوار الأرض، يسمى القمر. القمر هو أقرب الاجرام الفضاء الى الأرض ولكنه يختلف عن الأرض. ما هو وجه الاختلاف عن الأرض
14	User	The moon does not have plants or oxygen.	القمر ليس له غلاف جوي لذلك يختلف عن الأرض
15	LANA	<i>Well done</i> , the moon doesn't have atmosphere and water whereas the earth has.	<i>احسنت بارك الله فيك</i> • يختلف القمر عن الأرض بأن ليس له غلاف جوي كما أنه خال من الماء.

		Although the moon can appear a very bright white, its surface is actually dark. So from where the moon reflects the light?	القمر لا يضيئ بنفسه فهو انعكاس لضوء آخر. ما هو؟
16	User	The sun	الشمس
17	LANA	<i>Well don</i> , the moon reflects the sun light. we see the moon in different shapes during the month. What is the reason for that?	<i>احسنت بارك الله فيك</i> القمر لا يضيئ بنفسه فهو انعكاس ضوء الشمس نرى القمر في اشكال مختلفة خلال الشهر. فما سبب هذا الاختلاف؟
18	User	Because the moon rotates around the earth	بسبب دوران القمر حول الأرض
19	LANA	<i>Well done</i> , we see the moon in different shapes during the month because of the moon's rotations around the earth. How long does the moon take to rotate around the earth?	<i>احسنت بارك الله فيك</i> نرى القمر في اشكال مختلفة خلال الشهر بسبب دوران القمر حول الأرض ماهي المدة التي يتم فيها القمر دورته حول الأرض؟
20	User	One month	في شهر واحد
21	LANA	<i>Well done</i> , the moon takes 29 days (one month) to complete one rotation around the earth. The bright part in the moon is changing during the moon's rotation around the earth and this changing is called moon's phases. How many phases does the moon have?	<i>احسنت بارك الله فيك</i> يتم القمر دورته حول الأرض في 29 يوما أي شهر كامل يتغير شكل الجزء المضاء من القمر في أثناء دورانه حول الأرض فنراه في أشكال عديدة تسمى أطوار. فكم عدد هذه الاطوار؟
22	User	eight	ثمانية
23	LANA	<i>Well done</i> , the moon has eight phases. Tell me about another natural phenomenon, which occurs to the sun or the moon. Tell me what is it?	<i>احسنت بارك الله فيك</i> يتغير شكل الجزء المضاء من القمر في أثناء دورانه حول الأرض فنراه في ثمانية أطوار ظاهرتان تحدث للشمس والقمر وهما من آيات الله . اذكر لي اسمهما؟
24	User	I don't know	مادري
25	LANA	<i>Thanks for the trying</i> , the answer is eclipses. Eclipses can only occur when the Sun, Earth, and Moon are all in a straight line. So watch how the eclipse occurs. Solar eclipses occur at new moon, when the Moon is between the Sun and Earth. Using the earth model place the moon between the light and the earth. In contrast, lunar eclipses occur at full moon, when Earth is between the Sun and Moon. Using the earth model place the earth between the light and the moon.	<i>شكرا للمحاولة ولكن الأجابة المثلثي هي الخسوف والكسوف هما من آيات الله عز وجل. خسوف القمر يحدث عندما تقع الأرض بين الشمس والقمر فيمر القمر في منطقة ظل الأرض فيبدو لنا معتما. كسوف الشمس يحدث عندما يقع القمر بين الشمس والأرض فيحجب القمر ضوء الشمس عن الأرض</i> <i>استخدم نموذج الكرة الأرضية واجعل القمر بين المصباح والأرض). هذا يمثل كسوف الشمس) والآن استخدم نموذج الكرة الأرضية واجعل الأرض بين المصباح والقمر) هذا يمثل خسوف القمر)</i> شكرا لك واتمنى ان تكون استفدت من الدرس
26	User	Good bye	مع السلامة

Table 5.18 illustrated how the tutorial was given to the participants when their preferred learning style was kinaesthetic. The blue sentences (also in Italic) are the system feedback based on the participant's answer. The red sentences give instructions to the participant to interact with the models (the solar system model and the earth model). Figure 12 shows a picture of the models given to the participants.

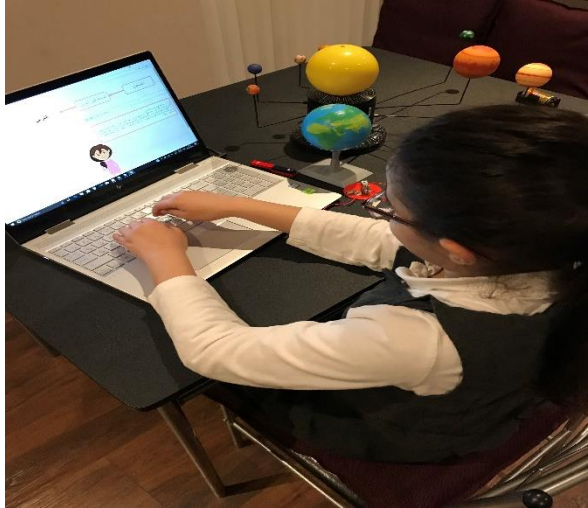


Figure 5.2: The solar system model and earth model

The example conversation also demonstrates the LANA-I engine and how it was able to recognise that utterance 14 from Table 5.18:

القمر ليس له غلاف جوي لذلك يختلف عن الأرض (The moon does not have atmosphere so it is different from the earth) was a variation of the scripted pattern: يختلف القمر عن الأرض بأن ليس له غلاف جوي (The moon is different from the earth that it does not have an atmosphere) and correctly respond to the user. The results of the log file analysis from LANA-I is illustrated in Table 5.19.

Table 5.19: Data analysis from the log file

CONVERSATION ANALYSIS	
METRIC	
Total number of conversations	24
Total number of utterances in all conversations	313
Average number of words per user utterance	3.1
Average number of utterances per conversation	13
Number of correct response from LANA	89.8%
Number of incorrect response from LANA	10.2%
Number of unrecognized utterances	38%
Number of utterances using Similarity Strength	57.18%
Number of rules fired	170
Average conversation duration (mins)	11.5
Percentage of conversations leading to goal	95.8%

In addition, the frequency analysis of the questionnaire results illustrated in Table 5.20 shows that all participants (N=24) satisfied with the usability of LANA-I. The results indicated that the majority (79.16%) of the participants felt that learning using (LANAI) is better than learning from the teacher in the class, meaning that **HB₁ can be accepted.**

Table 5.20: Frequency analysis LANA questionnaire

QUESTIONNAIRE RESULTS	Happy	Normal	Sad
I think LANA Tutor is enjoyable to use	100%	0	0
I think LANA had a right amount of information in one Tutorial.	91.6%	8.4%	0
I like the design and colour of LANA.	91.6%	8.4%	0
I think I learned useful information from LANA	91.6%	8.4%	0
I think I would like to use LANA again with other lessons.	91.6%	8.4%	0
I think talking with LANA-Is like as talking with my friend.	45.8%	50%	4.2%
I think I can learn from LANA better than the teacher in class	79.16%	20.84%	0

An additional finding from the results indicated that the majority (91.6%) of the participants said that they would use LANA-I again with other lessons, however less than half of the participant felt that talking with LANA-I was like as talking with their friend.

5.10 Discussion

The results illustrated from the first experiment that there was statistically significant difference (as shown in Table 5.10) between the user's learning gain with and without adapting to VAK learning style. The mean rank of the differences between the pre-test and post-test scores in each case are without adapting to the VAK ($M = 9.50$) and with adapting to the VAK ($M = 15.50$). The P-value of the mean's variation between the two groups is 0.03 (less than 0.05). In terms of the relative learning gain, the results indicated that there was statistically significant difference (as shown in table 5.11) between the user's relative learning gain with and without adapting to VAK learning style, p-value less than 0.05 ($P = .015$). This means that the users performed better when the tutorial is adapted to their preferred VAK learning style. In addition, the users who learned by adapting to VAK model were much happier with remembering what they have learned from LANA-I (91.7%) than users who learned without adapting to VAK model (50%). These results indicated that LANA-I can be adapted to the user's learning style and enhance their learning. However, the results show in Table 5.16 that users with Auditory learning style have not performed well compared with Kinaesthetic and Visual learning style users. This could be addressed by adding more audio materials in the system. Based on these findings HA_1 can be accepted in all (3 out of 3) metrics (Table 5.2 and table 5.3).

The results in Table 5.19 reveal that in general LANA-I performed well as an Arabic CA based on the number of correct and incorrect responses. The results show from the total conversations there were 281 correct responses (about 89.8 % from the total utterances) and 32 incorrect responses (about 10.2% from the total utterances). Moreover, from the conversations that did reach the end of the tutorial, the majority 95.8% of them reached the end, meaning the user stayed with the tutorial and did not interrupt the conversation before the end of the tutorial. It could be concluded from these results that the users did enjoy using LANA-I as a tutoring system, this also reflected in the opinion of the participants in the end user questionnaire (Table 5.20) who thought that LANA-I was enjoyable (100%) and a majority stated they would use the system again (91.6%).

An additional insight provided from these results was that the LANA-I engine is allowing the reduction of scripted patterns. The results illustrated that (57.18%) of all the utterances input by the users were actually different from the scripted patterns and in this case the system used the Short Text Similarity algorithm (Cosine algorithm). The log file shows that 34 unique utterances relating to 15 different rules were correctly recognised and dealt with by the LANA-I algorithm by firing the appropriate rule. Based on these findings HB₁ can be accepted as 10 out of 12 metrics tested were significant (Table 5.2 and Table 5.3).

However, the results show some of the weaknesses in LANA-I architecture, mainly the number of unrecognised utterances. LANA-I failed to recognise some utterances (38%) from the users. Upon further analysis of the log file it was found that some of these unrecognised utterances were due to:

- Spelling mistakes in users utterances. These mistakes are due to the phonological similarity of some of its alphabet characters that caused the engine to fail to recognise the sentence. This means that the strength of that utterance, was below the acceptable threshold set within LANA-I engine. For example, writing (اربعه) instead of (أربعة), which is the correct spelling. In this example, the two words have the same meaning but the first word has a common spelling mistake where each of the letters “أ” and “ة” has a different form in the two words.
- Word structure: some words have the same meaning but they are different because of the prefix and suffix. For example, the texts (الفصول الاربعة) and (اربع)

(ال) have the same meaning but they are different because of the prefix (ال) and the suffix (ه).

- The other cause for the unrecognised utterances was due to gaps in the knowledge base, which can be easily addressed by expanding the knowledge base. However, further research is needed to overcome the spelling and word structure issues.

5.11 Chapter Summary

This chapter has outlined and detailed the experimental methodology that used to evaluate the LANA-I and illustrated the results to validate LANA-I architecture proposed in Chapter 4. Two experiments have been conducted using the sample of 24 neurotypical participants (aged 10-12 years old) and their first language is Arabic. The success of (LANA-I) was evaluated by measuring the success of two hypotheses:

- (LANA-I) using VAK learning style can be adapted to the student learning style.
- LANA CA is an effective Arabic CA.

The results from the experiments, using a combination of subjective and objective measures, show a significant evidence to support both of the hypotheses. However, further research is needed to overcome the weaknesses, which is detailed in the next chapter.

The main novel contributions in this chapter are:

- A new general evaluation framework has been developed and tested from the objective and subjective perspective.
- Evaluation of (LANA-I) through two empirical evaluation experiments which validate the architecture components and the VAK learning style model with neurotypical students in Arabic primary schools.
- A discussion of the four research questions and the main findings was:
 - Using VAK learning style with LANA-I improves the participants (neurotypical) learning.
 - LANA-I can be adapted to the user's learning style and enhance their learning.

- LANA-I engine managed to reduce the number of scripted patterns by an average of 11% using the STS algorithm comparing to PM approach.
- LANA-I engine managed to respond correctly to the user by an average of 89%
- LANA-I is able to lead the user towards the goal of the tutorial, with 95.8% success of completing the tutorial.

Chapter 6 - LANA-II with Improved Architecture

6.1 Overview

This chapter aims to describe the development of the second prototype known as LANA-II which was designed in order to address the issues and the weaknesses based on the findings from the first evaluation (Chapter 5). The evaluation experiments revealed some positive results for the architecture and components of LANA-I, however, there were a number of areas highlighted that could be improved. The issues that were highlighted in the first evaluation were mainly due to grammatical features and the morphological nature of the Arabic language. Other issues that were identified were the need to further expand the knowledge base in order to increase the naturalness of the conversation.

The weaknesses that were made apparent through the end user evaluation and the log file are as follows:

- Arabic Language Features: Common spelling mistakes.
- Architecture Features: Cosine algorithm similarity calculation (Unrecognised words with prefix and suffix).
- Gaps in the knowledge base: Unrecognised utterances.

As well as addressing the weaknesses in the first prototype system, this chapter details the development of the TEACCH method (Section 6.5) that are used to educate autistic children and how this method could be effectively implemented into LANA-II.

The research and development decisions made in order to mitigate these weaknesses are as follows:

- Common spelling mistakes: To overcome the issue of common spelling mistakes and variations, a pre-processing feature was added to LANA-II architecture (see section 6.2).
- Cosine algorithm similarity calculation: The Cosine algorithm was improved in order to recognise and better deal with the words that have prefix, suffix or differences in the word composition (see section 6.3).
- Knowledge base expansion: The knowledge base was expanded in order for LANA-II to seem more natural during conversation. In addition, another

tutorial was added as well as general knowledge items were added to the database (see section 7.4).

- User Interface using TEACCH method: to make LANA-II suitable for Autistic children, TEACCH method was applied in the user interface (see section 7.5).

These improved components are detailed in the subsequent sections of this chapter. Experiments in chapter 7 will assess whether these changes contribute to improve the effectiveness and accuracy of LANA-II engine.

6.2 Morphologic Pre-processing

As mentioned in the previous section, one of the issues that needed to be addressed is the common spelling variations and mistakes made by user. These mistakes are due to the phonological similarity of some of Arabic alphabet characters. In Arabic language, the letters change forms according to their position in the word (beginning, middle, end and separate). Some of these letters undergo a morphological modification in writing which does not influence the meaning of the word. For example, the letter “ا” at the beginning of a word can take different forms: “أ”, “إ”, “آ” or “ا”. A typical example is found in Arabic word (اربعه) instead of (أربعة), which is the correct spelling. In this example, the two words have the same meaning but the first word has a common spelling mistake where each of the letters “أ” and “ة” has a different form in the two words. In order to overcome the problem of the variation in the representation of Arabic letters, pre-processing methods needed to be applied to the utterance before applying the similarity algorithm:

- Replacing أ , إ , آ by ا
- Replacing ي , ئ , ء by ى at the end of the words.
- Replacing ة by ه at the end of the words.

The pre-processing method is applied by the controller (Chapter 4). The controller will be responsible for:

- Filtering the utterance to remove the special characters (i.e. \$, &, *, !, ?, “”, £, (), ^)
- Checking for rude words.
- Applying the pre-processing method.

6.3 Cosine similarity

Cosine similarity (Qian et al., 2004) was used in the LANA-I algorithm in order to calculate the string similarity between two texts. It determines the similarity between two pieces of text by representing each piece of text in the form of word vector. A word vector is a vector of length N where N is the number of different tokens in the text. The similarity is computed as the angle between the word-vectors for the two sentences in vector space. This means that the cosine compares the two text using the word level approach. For example, the texts (الفصول الاربعه) and (اربع فصول) have the same meaning but they are different because of the prefix (ال) and the suffix (هـ). In this example, the similarity must be (1) but if the cosine algorithm is applied, the similarity results will be zero, which means there is no similarity between the two texts.

Therefore, in order to improve the similarity calculation, a novel algorithm was implemented. The cosine similarity was adjusted to use a character level approach. This means that each word will be compared with other word character by character after applying the Bi-gram method (Kondrak, 2005). The general process of the enhanced cosine algorithm is illustrated in the following steps for the two utterances t_1 and t_2 :

t_1 : الفصول الاربعه (four seasons with the prefix (ال) and the suffix (هـ))

t_2 : اربع فصول (four seasons without the prefix and the suffix)

Step 1: Create Matrix[][] where the columns are the unique words from t_1 and t_2 , and the rows are the words sequence of t_1 and t_2 .

Step 2: Calculate the similarity between each word by applying the bi-gram method. It is difficult to compare the letters of the two words because some words have the same letters but different meaning. For example, the two words in English (rose, sore) or in Arabic (عنب, نبع) have identical letters but they are different in the meaning. In the bigram, each word is represented by a list of its constituent n -grams, where n is the number of adjacent characters in the substrings, $n=2$ in the bigram. Using the bigram list, similarity measures between pair of words are calculated based on cosine similarity approach. Table 6.1 shows an example of calculating the similarity of two utterances t_1 and t_2 .

Table 6.1: Calculate the similarity between pair of words

Matrix [][]=

فصول (seasons without prefix)	اربع (four without prefix and suffix)	الاربعه (four with the prefix (ال) and the suffix (هـ))	الفصول (seasons with the prefix (ال))		
0.77	0	0.2	1	الفصول (seasons with the prefix (ال))	t ₁
0	0.71	1	0.2	الاربعه (four with the prefix (ال) and the suffix (هـ))	
0	1	0.71	0	اربع (four without prefix and suffix)	t ₂
1	0	0	0.77	فصول (seasons without prefix and suffix)	

For example, the words ‘فصول’ and ‘الفصول’ have the same meaning but they are different because the second word has the prefix ‘ال’. If the cosine similarity is applied directly the similarity between these words will be zero. In the new algorithm, the bigram is applied first between the words, in this example the bigram will be:

Word1: فصل, صور, ول

Word 2: ال, لف, فصل, صور, ول

Then the cosine similarity is applied between these words:

Union: ال, لف, فصل, صور, ول

Vector1: 1,1,1,0,0

Vector2: 1,1,1,1,1

The cosine similarity is calculated using (Equation 1, Chapter 4), where t₁ is vector1 and t₂ is vector 2. The similarity between the two words is 0.77. Table 6.1 shows the similarity measures between each two words after applying the bigram method and the cosine similarity.

Step 3: The last step here is to write the vector for each string when the similarity between the words is greater than 0.40:

$$\begin{array}{ccccc} t_1 & 1 & 1 & 0.71 & 0.77 \\ t_2 & 0.77 & 0.71 & 1 & 1 \end{array}$$

Then the Cosine similarity measure is applied ((Equation 1, Chapter 4): $SIM(t_1, t_2) = 0.96$). Table 6.2 shows a pseudo-code representation of the new similarity algorithm.

Table 6.2: a pseudo-code representation of the new similarity algorithm.

```
String1= Words sequence of Pattern
String2= Words sequence of user's Utterance
Union= all words (String1 AND String2)
Unique = Unique words (String1 AND String2)
Bigram_word= each word is represented by a list of its constituent n-grams, where n=2 character

Calculate the similarity between pair of words
FOR EACH (Union) AND (Unique)
{
    Bigram_word1= Bigram (Union)
    Bigram_word2= Bigram (Unique)
    Matrix [Union][Unique]= Cosine Similarity ((Bigram_word1),( Bigram_word2))
}

Calculate the String1 vector
FOR EACH (Unique) AND (String1)
{
    IF (Matrix [String1][Unique]>=0.40)
        String1_vectors =Matrix [String1][Unique]
    ELSE
        String1_vectors =0
}

Calculate the String2 vector
FOR EACH (Unique) AND (String2)
{
    IF (IF (Matrix [String2][Unique]>=0.40)
```

```

String2_vectors = Matrix [String2][Unique]
ELSE
String2_vectors =0
}
COSINE SIMILARITY (String1_vectors, String2_vectors)
IF COSINE SIMILARITY >=80 THEN
Similarity=TRUE
ELSE
Similarity=FALSE

```

The new LANA-II similarity algorithm explained above, solves the complex word order issue that comes with the Arabic language. It also significantly reduces the number of scripts that have to be scripted to deal with the issue of word order or word structure (adding the prefix or suffix). An example of this is illustrated in Table 6.3, where the scripted pattern will cover five different patterns.

Table 6.3: The number of scripts patterns for the pattern (Four seasons)

Scripted pattern	الأربعة (four with the prefix (ال) and the suffix (هـ))	الفصول (seasons with the prefix (ال))
Patterns covered	فصول (seasons without prefix)	اربع (four without prefix and suffix)
	اربع (four without prefix and suffix)	فصول (seasons without prefix)
	الأربعة (four with the prefix (ال) and the suffix (هـ))	فصول (seasons without prefix)
	فصول (seasons without prefix)	الأربعة (four with the prefix (ال))
	الأربعة (four with the prefix (ال))	الفصول (seasons with the prefix (ال))

Table 6.4 shows the differences in the similarity scores between LANA-I and LANA-II similarity algorithm. As shown in the table, it is noticed that LANA-II performed better when the two strings are different because of the word structure such as the prefix (line 1) and spelling mistakes such as duplicate the letter (line 7). In addition,

LANA-II performed better when there are differences in the word composition such as the words **اختلاف** and **يختلف** (difference, differ) noun and verb (line 2).

Table 6.4: Comparison of the similarity algorithm scores between LANA-I and LANA-II

Line number	Reason	LANA-I	LANA-II	Utterance	Pattern	Question
1	Prefix and suffix	0	0.98	الأربع فصول	الفصول * الأربعة	تدور الأرض حول الشمس كل سنة. ماهي نتيجة هذا الدوران؟
2	Word composition and prefix	0	0.90	اختلاف الفصول السنة	تختلف فصول السنة	لقمر هو أقرب الاجرام الفضاء الى الأرض ولكنه يختلف عن الأرض. ما هو وجه الاختلاف عن الأرض
3	prefix	0.23	0.89	ليس للقمر غلاف جوي	القمر ليس له الغلاف الجوي	تستغرق الأرض لتدور حول محورها 24 ساعة أي يوم كامل ماذا ينتج عن هذا الدوران؟
4	Word composition and prefix	0.50	0.91	يختلف بعدم وجود هواء	الاختلاف بعدم وجود الهواء	
5	prefix	0.55	0.82	القمر لا يوجد به الأكسجين	لا يوجد به غاز الأكسجين	
6	prefix	0	0.92	انهار والليل	الليل والنهار *	
7	Duplicate the letter	0.52	1	الليل والنهار	الليل والنهار *	
8	prefix	0	0.98	كواكب	الكواكب *	

6.4 Knowledge base expansion

The knowledge base was expanded to address the finding from the end user questionnaire and the log file. The participants expressed that the naturalness in LANA-I conversation is low. In addition, the results from the LANA-I evaluation (Chapter 5) found that there are 38% unrecognised utterances. This means there are some utterances that LANA-I could not recognise because of the gap in the general and tutorial knowledge base. Therefore, the knowledge base was expanded to make LANA-II more natural in term of conversation.

The domain specific knowledge was increased by adding new contexts. This process involved further interviews with three primary school teachers in Saudi Arabia, who teach the science subject (see Appendix G for interview questions). The result of the interviews recommended to expand the knowledge base to include another tutorial related to the science topics. In addition, the first tutorial was expanded by adding new patterns. All unrecognised utterances resulting from weakness in the knowledge base from the LANA-I evaluation were added as new patterns to the knowledge base. Furthermore, the general topics of the knowledge base was expanded to include more topics and patterns such as greeting, weather and sport.

6.4.1 Tutorial 2: Electrical circuit and the magnetism

A new tutorial was added in LANA-II. After it was designed and approved by the teachers from Saudi Arabia, the tutorial questions were mapped to the VAK learning style model. The pictures and the video clips that are used for the visual learner, were provided from the Saudi Ministry of education website (Education, 2017). All these materials were evaluated and approved by the Primary school teachers in Saudi Arabia. Table 6.5 shows how the tutorial questions for the second topic are mapped to the learning style model VAK. Figure 6.1 shows the models that are used with the Kinaesthetic learning style.

Table 6.5: The second Tutorial

Question 1: Electricity has a big impact in our lives; we use it to the light of the lamp. Electricity is generated from electric charges, which are very small and cannot be seen. Electric charges are divided into two types: negative charges and positive charges. When a charged object comes near to another object, they will either attract or repel each other. What happens when similar charges converge (positive or negative)? Answer: If the charges are opposite - they attract	V: video A: Audio K: Electrical circuit.
Question2: If we rub the balloon on the woolly object and then hold the balloon up on the wall. What will happen? Answer: The balloon stay stuck on the wall.	V: Picture A: Audio K: Electrical circuit.
Question 3: Electrical charges can be passed through some materials to be power supply. The power supply only works in a closed path. What is the name of this path? Answer: Electrical circuit.	V: Picture A: Audio K: Electrical circuit.
Question 4: The circuit consists of three basic parts. what is it? Answer: energy source, an electrical load (device), a conductor (wire).	V: Picture A: Audio K: Electrical circuit.
Question 5: The power is applied when the circuit is closed which means it is connected. How can we control the flow of electricity in the circuit to switch on or off the lamp? Answer: By connecting an electric switch.	V: video A: Audio K: Electrical circuit.
Question 6: Magnet is a material used to attract certain objects. Do you know what objects are attracted by the magnet? Answer: Objects made of iron, nickel or cobalt.	V: Picture A: Audio K: Magnet.
Question 7: Magnet can pull or push some objects without touching it, because there is an area around the magnet showing the effect of its power. What is this area called? Answer: This area is called the magnetic field.	V: Picture A: Audio K: Magnet.
Question 8: The magnet consists of two poles. What are they? Answer: The north pole and the south pole.	V: video A: Audio K: Magnet.
Question 9: What happens when we if you bring two north poles together, or two south poles together in two magnets? Answer: They repel and the magnets push each other away.	V: video A: Audio K: Magnet.
Question 10: What happens when we bring a north pole and a south pole together in two magnets? Answer: They attract and the magnets may stick together.	V: video A: Audio K: Magnet.

Each question in the tutorial is mapped to Visual learning style through pictures and videos, Auditory learning style through audios, Kinaesthetic learning style through models and instructions.

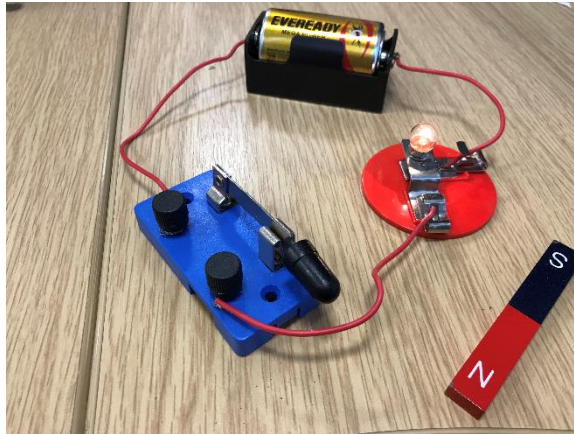


Figure 6.1: The models that are used with the Kinaesthetic learning style.

6.5 Mapping TEACCH method to the tutorial

TEACCH (the Treatment and Education of Autistic and related Communication handicapped Children) (Mesibov et al., 2003) is an education programs approach for people with ASD. This approach is developed by Eric Schopler (Schopler and Reichler, 1972) at the University of North Carolina in 1972. TEACCH is based on Structured Teaching technique for autistic students. This program is used in UK schools since 1990 then the National Autistic Society start collaboration with the Division TEACCH.

The aim of structured teaching is to enhance the teaching methods and the environment by considering the cognitive ability, needs and interest of the autistic students. In addition, the structured teaching aims to organize the classroom and to make the teaching methods and style friendly for autism. It is reported that pupils with ASD who used this approach, become calmer, and can work productively and independently for longer periods of time. Many advantages could be gain from using the Structured Teaching, as a method of delivering the curriculum. Such as, enhancing and facilitating the teaching and learning process. In addition, it can improve access to the curriculum for many pupils with ASD (Mesibov et al., 2003). TEACCH has four main components: physical structure, daily schedules, work system, and visual cues and instruction.

6.5.1 Physical Structure

The Physical structure is to arrange the materials, furniture and general surrounding in order to enhance the environment. In LANA-II, the physical structure is applied by enhancing the Graphics User Interface (GUI) making the screen less complicated as possible by adding only relevant material that is mandatory for that page. In addition, the colours were chosen to be suitable to the autistic students by using soft and mild colours (Pavlov, 2014).

6.5.2 Work System

This work system aims to organise the teaching activity by explaining the activity, the required time to finish this activity, undertake evaluation the student's progress and what will they gain after completing the activity. In LANA-II, the character LANA will provide the student an explanation of each interface. For example, the character LANA in Figure 6.2 explained the pre-test and why and how to do it.



Figure 6.2: applying the work system approach of TEACCH method

6.5.3 Daily Schedule

This provides all activities and their order, which will be conducted during the day. This helps the autistic students to understand clearly, what they have to do and give a predictable environment as they need (Mesibov, 2018). In LANA-II, an activity timeline in each interface is added that shows the current activity (Figure 6.3 and Figure 6.4).

Figure 6.3: Applying the scheduling approach of TEACCH method

Figure 6.4: The activity timeline in the pre-test interface..

6.5.4 Visual Cues and Instruction

Visual structure is used to clarify the tasks; it provides the pupil with ASD with information on how to complete the task. It is reported that the visual skills is usually strong in autistic pupils (Banire et al., 2015; O'Connell et al., 2013; Snyder et al., 2012). Visual information can be very helpful in developing understanding for pupil with ASD. Visual information has three kinds, which are utilised within LANA-II. The first type is the visual clarity, which engages the autistic pupil attention to important information. In LANA-II the conversation is coloured in which the red

colour is the question, the black colour is the user utterance, and the blue colour is the instruction (see Figure 6.5).



Figure 6.5: Visual instruction approach of the TEACCH method.

The second type is the visual organisation, which is the way to organise the tasks to allow the pupil with ASD focus on the task and complete it. In LANA-II, the using of the activity timeline (Figure 6.3, Figure 6.4, and Figure 6.5) helped to organise the tasks and helped the pupil with ASD to complete the whole tutorial.

The third type is the visual instruction, which are written cues that gives the pupils with ASD information on what is the next task and how to complete it. In LANA-II, the instruction interface is used before each activity in order to give the pupil information about the next task (see Figure 6.2).

6.6 Updated Architecture Diagram

Figure 6.6 illustrates the updated architecture from the LANA-I architecture (Chapter 4). The figure outlines all the new components presented in this chapter and how they interact with each other to overcome the issues highlighted through the first end user evaluation.

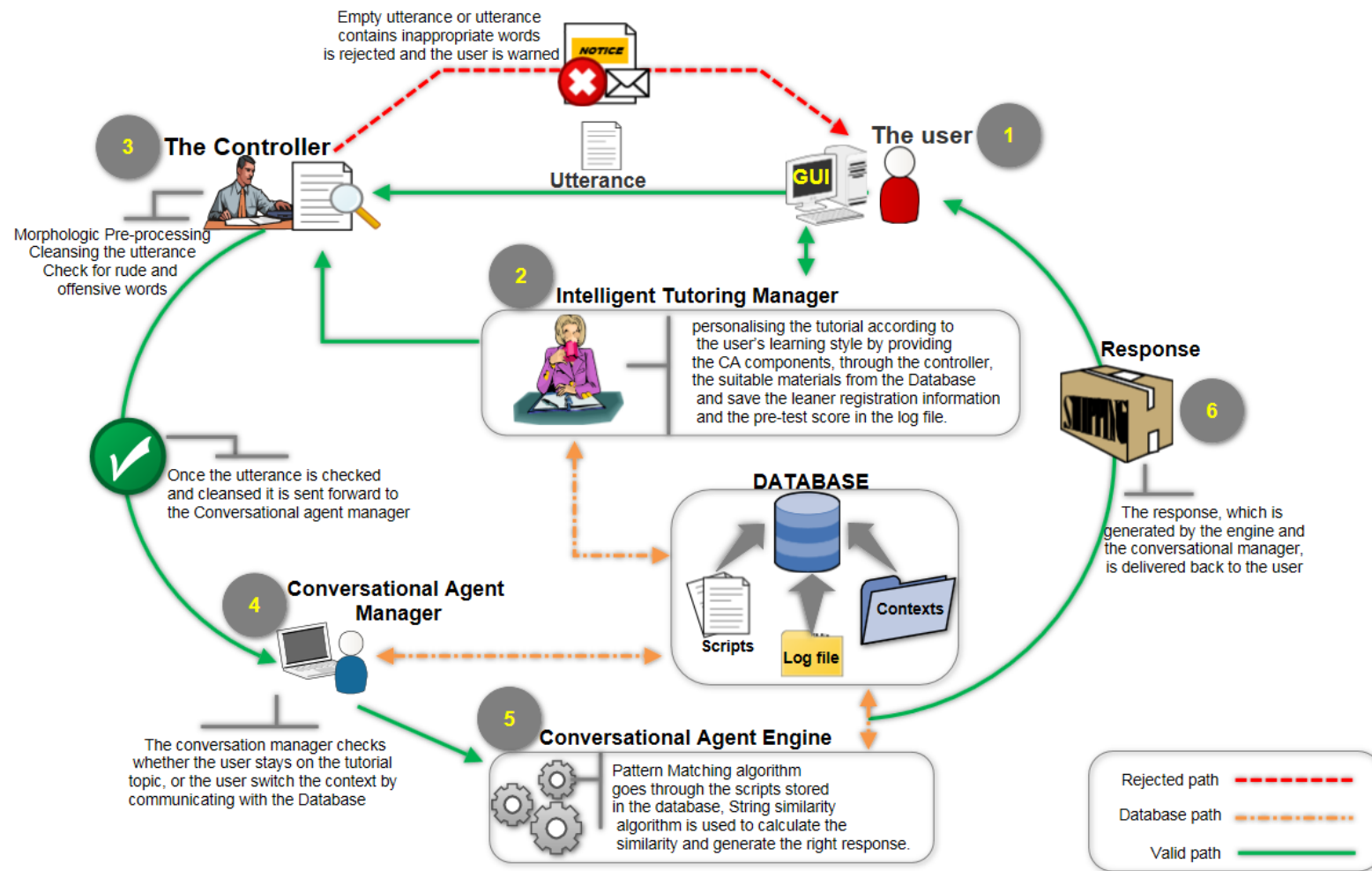


Figure 6.6: LANA-II Architecture and the Utterance Processing

6.7 Scripting tool

Writing the patterns and storing them in the database is time consuming. Therefore, a scripting tool has been developed as illustrated in Figure 6.7. The aim of developing this tool is to make the task of scripting easier. The scripting tool is not connected to the engine, it is connected to the LANA-II Database. The tool acts an interface between the scripter and the knowledge base/database.

The image shows a web-based interface titled "Scripting Tool". It features a light blue background. On the left side, there are four labels: "PatternID", "ContextNumber", "RuleNumber", and "PatternContent". To the right of these labels are four corresponding white input fields with blue borders. Below the input fields is a small, rectangular "Submit" button with a light orange background and a thin black border.

Figure 6.7: The scripting tool.

6.8 Chapter Summary

This chapter has outlined the additional research, development and approaches undertaken to address the weaknesses brought to light during the evaluation of (LANA-I). TEACCH method has been added to the GUI in order to adapt the tutorial environment to the autistic students learning. Additionally, new string similarity algorithm and pre-processing method have been developed and added to architecture the in order to improve the robustness and accuracy of LANAs-II engine. The knowledge base was expanded by adding a new tutorial and expanding the general topics in order to address the unrecognised utterances from (LANA-I) evaluation (Chapter 5). The new updated architecture of LANA-II will undergo end user evaluation with the intention of gauging whether or not the new components have any impact on the success and effectiveness of LANA-II as a CITS compared to LANA-I. The evaluation methodology and results are outlined in the following chapter.

The main novel contributions in this chapter are:

- Presentation of a new architecture (Improved from LANA-I) for creating an Arabic CITS.
- A new LANA-II Conversation Agent (CA) algorithm that reduces the scripting effort and unrecognised utterances by processing the spelling mistakes and morphological nature of the Arabic language, which were issues in LANA-I.
- Cosine algorithm similarity calculation: the similarity algorithm (used to match inputs with responses in LANA I) was improved in order to recognise and better deal with the words that have prefix, suffix or differences in the word composition.
- Knowledge base expansion: The knowledge base was expanded in order for LANA II to seem more natural during conversation. In addition, more domain specific (i.e. Science topics) as well as general knowledge were added to the database.
- Development a new CITS called LANA II, which adapts the TEACCH method that is used to educate autistic children. LANA II is the first CITS that mapped TEACCH method to the user interface.

Chapter 7 - LANA-II Evaluation Methodology and Results

7.1 Overview

This chapter presents the design of an evaluation methodology for the enhanced component of LANA-I architecture, which was proposed in Chapter 4. The LANA-I evaluation (Chapter 5) was focused on evaluating metrics related to the different components of LANA-I architecture. In the LANA-II evaluation, these metrics will be carried over to evaluate the effectiveness of the enhancements made to those components. The metrics from LANA-I will be compared with the metrics from the LANA-II. The intention behind this is to bring to light any significant improvements between the two systems revealed by the metrics. In addition, the LANA-II evaluation will test the adaptation of the VAK model to the autistic students learning style. The following sections outline the evaluation methodology through three experiments, the aim of the first experiment is to test two questions related to Hypothesis A (LANA-II can be adapted to the autistic students learning style), which are:

Question1: Can LANA-II improve the autistic students' perceptions of the learning experience by adapting to VAK learning Style?

Question2: Can LANA-II by adapting to VAK learning style improve the learning gain of the autistic students?

This experiment was conducted with the autistic students using LANA-II in two tutorials (the solar system, Chapter 4) and (the electricity and magnetic, Chapter 6). The second experiment was conducted to test the Hypothesis B (The enhancements made to LANA-II architecture improve the overall effectiveness of LANA-II engine):

Question 1: Is LANA-II engine more effective than (LANA-I) engine?

Question 2: Do the improvements added in the LANA-II engine improve perceptions over (LANA-I) by the user?

This experiment was conducted with the neurotypical students using LANA-II in the solar system tutorial as this tutorial was used in LANA-I evaluation (see the tutorial in chapter 4).

The third experiment will be a case study to test and evaluate LANA-II qualitatively with three autistic students. The aim of the third experiment is to test a question related to Hypothesis C (LANA-II is appropriate as a virtual tutor for Autistic students).

Question: Are the autistic students satisfied with the usability of LANA-II?

The results derived from the three experiments will contribute towards concluding the main research questions, which are:

1. Can an Arabic Conversational Intelligent Tutoring System (CITS) be adapted to the learning styles for autistic children?
2. Can an Arabic Conversational Intelligent Tutoring System (CITS) for children with autism enhance their learning?

7.2 Experimental Design

This section describes the experimental design for evaluation of LANA-II. The first experiment was conducted with Autistic children on the target age group (10-12) years old whose first language is Arabic and have a high functioning autism (HFA) who have no language problems and intellectual impairment. This experiment was conducted to measure the capabilities from different aspects such as tutoring success, VAK learning style model and user evaluation. The second experiment was conducted with participants in the target age group (10-12) years old whose first language is Arabic. None of the children have been diagnosed as Autistic. This experiment was test the enhancement made in LANA-II engine to compare the metrics from the LANA-I with the same metrics from LANA-II. This will bring to light whether the enhancements made in LANA-II components improve LANAs conversation. The third experiment was a case study with three autistic students to test if LANA-II is appropriate as a virtual tutor for Autistic students. The results derived from the three experiments will contribute towards concluding the main research question.

7.2.1 Ethical considerations

The following ethical guidelines were put into place for the research period (see Appendix D). MMU ethical approval was obtained (No. SE151615, Appendix L):

- No personal information about any child will be stored.

- Parents who do not wish their children to be included in the project and do not sign the consent form will be excluded.
- The children's teachers will be present as observers, with the authority to withdraw them from the study if they see fit.

7.2.2 Hypothesis

The main hypotheses of the experiments are:

- **HA₀**: LANA-II using the VAK model cannot be adapted to the Autistic learning style.
- **HA₁**: LANA-II using the VAK model can be adapted to the Autistic learning style:

This hypothesis requires a VAK questionnaire to be filled before starting the tutorial in order to identify the autistic student's learning style and then the student is automatically provided with suitable materials depending on their learning style result: video and images for visual learning style, sounds for auditory learning style, and models and touching screen for the kinaesthetic learning style.

- **HB₀**: the enhancement made to LANA-II architecture did not improve the overall effectiveness of LANA-II engine.
- **HB₁**: The enhancements made to the LANA-II architecture improve the overall effectiveness of LANA-II engine:

This hypothesis relates to the improvements and changes made to the LANA-II similarity algorithm. The newly proposed similarity algorithm has an impact on the accuracy and effectiveness of LANA-II engine and reduces the percentage of unrecognised utterances.

- **HC₀**: LANA-II is not appropriate as a virtual tutor for Autistic students.
- **HC₁**: LANA-II is appropriate as a virtual tutor for Autistic students.
This hypothesis will be measured using a case study with three autistic students including interview and observation.

7.2.3 Designing of evaluation metrics

In order to successfully test the hypotheses defined in section 7.2.2, a number of metrics are required. Figure 7.1 shows the GQM model that consists of three goals

related to hypotheses and each goal has questions. These questions use a combination of subjective metrics (questionnaire metrics), and objective metrics (log file metrics).

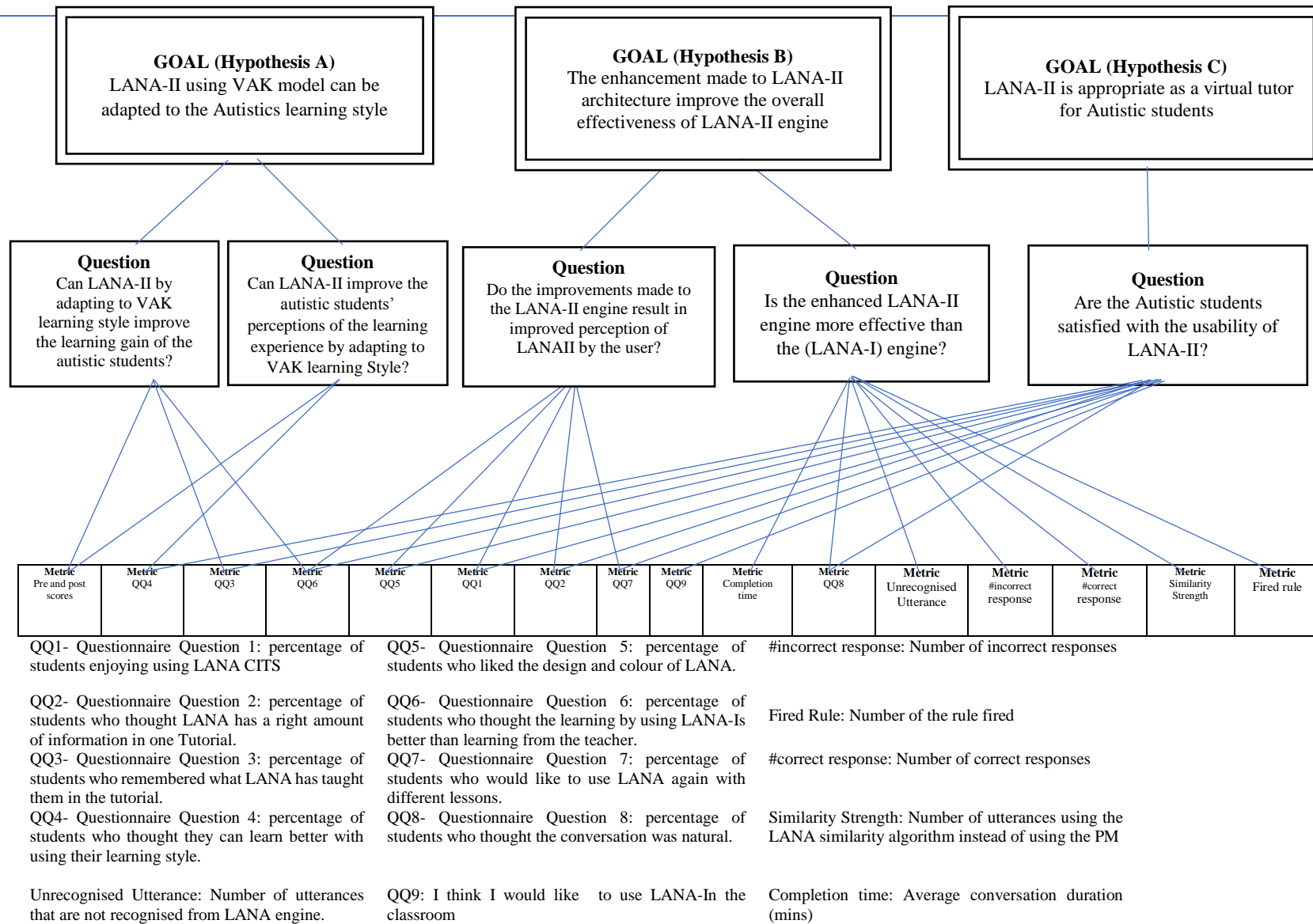





























Figure 7.1: GQM Model for the second prototype evaluation.

The feedback questionnaire that was used in the LANA-I evaluation (Chapter 5) has been modified in this experiment by adding one more question (Q9). This is to evaluate whether the autistic student is satisfied with LANA-II and if he/she would like to use it in their classroom. Table 7.1 shows the modified feedback questionnaire.

Table 7.1: The modified feedback Questionnaire (Translated version).

1. I think LANA Tutor is enjoyable to use			
2. I think LANA had a right amount of information in one Tutorial.			
3. I think I remember what LANA has taught me in the tutorial.			
4. I think I can learn better with using my learning style (Visual – Auditory -Kinaesthetic).			
5. I like the design and colour of LANA.			
6. I think I can learn from LANA better than my teacher.			
7. I think I would like to use LANA again with other lessons.			
8. I think talking with LANA is like talking with my friend.			
9. I would like to use LANA in the classroom			

The subjective metrics, the methods used with each metric and the related hypotheses are illustrated in Table 7.2. Table 7.3 shows the objective metrics and their evaluation methods and the related hypothesis.

Table 7.2: Subjective Evaluation Metrics

Subjective Metrics		
Metrics to be Evaluated	Method of Evaluation	Hypothesis
QQ1: LANA Tutor is enjoyable to use	Questionnaire	Hypothesis B Hypothesis C
QQ2: LANA has a right amount of information in one Tutorial.	Questionnaire	Hypothesis B Hypothesis C
QQ3: I remember what LANA has taught me in the tutorial.	Questionnaire	Hypothesis A Hypothesis C
QQ4: I can learn better with using my learning style.	Questionnaire	Hypothesis A Hypothesis C
QQ5: I like the design and colour of the interface.	Questionnaire	Hypothesis B Hypothesis C
QQ6: I think I can learn from LANA better than my teacher.	Questionnaire	Hypothesis A Hypothesis B Hypothesis C
QQ7: I think I would like to use LANA again with other lessons.	Questionnaire	Hypothesis B Hypothesis C
QQ8: The conversation was natural.	Questionnaire	Hypothesis B Hypothesis C
QQ9: I would like to use LANA in the classroom	Questionnaire	Hypothesis C

Table 7.3: Objective Evaluation Metrics

Objective Metrics		
Metrics to be Evaluated	Method of Evaluation	Hypothesis
Completion time	Log file	Hypothesis B
Number of unrecognised Utterance	Log file	Hypothesis B
Number of correct and incorrect response	Log file	Hypothesis B
Similarity Strength	Log file	Hypothesis B
The rule fired	Log file	Hypothesis B
Pre-test and Post-test	Log file	Hypothesis A

The hypothesis HA_1 will be evaluated through the differences between the pre-test and the post-test scores and through the end user questionnaire. The hypothesis HB_1 will be evaluated through a comparison between the metrics from the first set of data (from LANA-I -Chapter 6) and the metrics from second set of data (from LANA-II). This comparison will show if there is any statistically significant improvement between the two data sets. The hypothesis HC_1 will be evaluated through a case study that was conducted with three autistic students.

7.3 Experimental Methodology

This section will describe the experimental methodology in evaluating LANA-II. The tutorial that was given by LANA-II is based on Science tutorials for children age group

(10-12) years old (Chapter 6, section 6.4.1). All experiments were undertaken in a classroom environment, and each student individually worked on a computer. The computer used in this experiment is a touchscreen computer (Windows 10) in order to let the students use both the keyboard and the screen to interact with LANA-II. The teacher was present during the experiment and the students were given instructions about using LANA-II.

During the tutorial session, all the questions from the system, the student answers, and the conversation variables/metrics were captured and saved in the log file.

7.4 Experimental Data Analysis

In order to analyse the log file data, a normality test (Chapter 5, section 5.7) was conducted. Appendix H shows the histograms for the questionnaire results and the pre-test and post-test scores along with the results of the normality test. The histograms and the Shapiro-Wilk test shows that the data is not normally distributed. Therefore, non-parametric tests will be used to analyse the data, which are, Chi-Square test, Wilcoxon test and Mann-Whitney U test (see Chapter 5, section 5.7.1, section 5.7.2, section 5.7.3). For the effect size, the same equation will be used (Equation 2, chapter 5).

7.5 Experiment 1: Adapting the VAK model to the autistic students learning Style

7.5.1 Participant

(Aljameel et al., 2016) reviewed Nineteen articles of technology interventions for children with ASD. The average participants in these studies was 13. In LANA-II experiment, thirty autistic students were recruited from five mainstream primary schools, that provide care for children with ASD in Saudi Arabia and Bahrain. Each child had previously received a community clinical diagnosis of high functioning autism spectrum disorder or Asperger's, who they have no language problem and intellectual impairment. All participants with the age group (10-12) years old and their first language is Arabic and none of them had a previous experience using LANA-II. An information sheet about the project and its aims and a consent form were sent to participant's parents to get their permission before conducting the experiment (see Appendix D and Appendix E for the information sheet and consent form). The

participants were divided into two equal sized groups. The first group was the Control Group (15 participants) used the LANA-II without adaptation to their VAK learning style. The second group was the Experimental Group (15 participants) whom used LANA-II with adaptation to their VAK learning style.

7.5.2 Experiment 1 Methodology

The aim of this experiment was to test the hypothesis HA_1 based on the metrics outlined in section 7.2.3. The experiment was based on the pre-test and post-test scores with and without adapting to VAK learning style model. The subjective and objective metrics in Table 7.4 were used to answer the two questions related to Hypothesis A (LANA-II can be adapted to the autistic student learning style), which are:

Question1: Can LANA-II improve the autistic students' perceptions of the learning experience by adapting to VAK learning Style?

Question2: Can LANA-II by adapting to VAK learning style improve the learning gain of the autistic students?

Table 7.4: Experiment 1 metrics

Metrics to be Evaluated	Method of Evaluation	Hypothesis
Pre-test and post-test	Log file	Hypothesis A
QQ3: I remember what LANA has taught me in the tutorial.	Questionnaire	Hypothesis A
QQ4: I can learn better with using my learning style.	Questionnaire	Hypothesis A
QQ6: I can learn from LANA better than learning from my teacher.	Questionnaire	Hypothesis A

7.5.3 Experiment 1: Results and Discussion

This section will present the collected results of four metrics outlined in Table 7.4. The data was extracted from the tutoring log file (pre-test and post-test scores), and the questionnaire (Questions 3, 4 and 6). These two data (pre-test and post-test scores, and Questionnaire) were used to determine whether the hypothesis A should be accepted or rejected, which is: LANA-II can be adapted to the autistic student learning style.

7.5.3.1 Pre-test and Post-test – Results and Discussion

Table 7.5 and Table 7.6 show the results of Wilcoxon Signed Ranks test carried out to determine if there was any difference between the pre-test and post-test scores in the Experimental Group.

Table 7.5: the difference between the mean of the pre-test and post-test scores in the Experimental Group

Descriptive Statistics								
	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
Pre-test	15	4.2000	1.56753	1.00	7.00	3.0000	4.0000	5.0000
Post-test	15	9.8000	.41404	9.00	10.00	10.0000	10.0000	10.0000

Table 7.6: The P Value between the means from Table 7.5

Test Statistics ^a	
	Post-test - Pre-test
Z	-3.431 ^b
Asymp. Sig. (2-tailed)	.001
a. Wilcoxon Signed Ranks Test	
b. Based on negative ranks.	

When comparing the mean ranks in Table 7.5 of the two scores, the post-test scores were ranked higher, which means that the students performed better in the post-test. As Table 7.6 shows p-values less than 0.05 ($p = .001$), it can be concluded that there is a statistically significant difference between the pre-test and post-test scores in the Experimental Group. The effect size (r) is calculated using (Equation 2) and found that $r = 0.6$ indicating a large effect size using (Cohen, 1988) criteria.

Table 7.7 and Table 7.8 show the results of Wilcoxon Signed Ranks test carried out to determine if there was any difference between the pre-test and post-test scores in the Control Group.

Table 7.7: the difference between the mean of the pre and post-test scores in the Control Group

Descriptive Statistics								
	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
Pre-test	15	5.2000	1.01419	4.00	7.00	4.0000	5.0000	6.0000
Post-test	15	7.6000	.73679	6.00	9.00	7.0000	8.0000	8.0000

Table 7.8: The P Value between the means from Table 7.7

Test Statistics ^a	
	Post_test - Pre_test
Z	-3.497 ^b
Asymp. Sig. (2-tailed)	.0004

a. Wilcoxon Signed Ranks Test

b. Based on negative ranks.

The results from Table 7.7 and Table 7.8 show that there is statistically significant difference between the pre-test and post-test scores p-values less than 0.05 ($p = .0004$) and the mean value in the post-test is greater than the mean in the pre-test, which indicates that the students performed better in the post-test. The effect size (r) is calculated using the (Equation 2) and found that $r = 0.6$ indicating a large effect size using (Cohen, 1988) criteria.

Table 7.9 illustrates the results of the Mann-Whitney U test conducted in order to measure whether there was a statistically significant difference in the learning gain (Equation 3, chapter 5) between the two groups (Control Group: without using VAK, Experimental Group: with using VAK).

Table 7.9: Mann-Whitney U test for the differences between the pre and post scores between Group1 and Group2.

Ranks				
	Group	N	Mean Rank	Sum of Ranks
Learning gain	Control	15	8.50	127.50
	Experimental	15	22.50	337.50
	Total	30		
Test Statistics ^a				
		Learning gain		
Mann-Whitney U		7.500		
Wilcoxon W		127.500		
Z		-4.450		
Asymp. Sig. (2-tailed)		.000009		
a. Grouping Variable: VAK				

From this result, it can be concluded that the differences between the pre-test and post-test scores between Control Group and Experimental Group are statistically significant p-value less than 0.05 ($P = .000009$). The effect size (r) is calculated using the

(Equation 2) and found that $r = 0.8$ indicating a very large effect size using (Cohen, 1988) criteria.

When comparing the mean ranks it can be seen that the Experimental Group – that used VAK - had a higher rank than the Control Group, meaning that the learning gain for the Experimental Group was significantly improved compared to the learning gain of Control Group. This means that adapting the tutorial to VAK learning style, improve the students' learning gain.

Relative learning gain (A. Latham et al., 2014) was also measured in this experiment. (see Chapter 5, section 5.8). Table 7.10 illustrates the results of the Mann-Whitney U test conducted in order to measure the relative learning gain (Equation 4) between Control Group and Experimental Group.

Table 7.10: Mann-Whitney U test for the relative learning gain between Control Group and Experimental Group.

Experimental Group				
Ranks				
	Group	N	Mean Rank	Sum of Ranks
Relative Learning gain	Control	15	8.00	120.00
	Experimental	15	23.00	345.00
	Total	30		
Test Statistics ^a				
	Learning gain			
Mann-Whitney U	.000			
Wilcoxon W	120.000			
Z	-4.838			
Asymp. Sig. (2-tailed)	.000001			
a. Grouping Variable: VAK				

Table 7.10 shows that there is a difference in the mean value between Control Group and Experimental Group. The average improvement in test scores in the Experimental Group were improved more than the Control Group. The ranks in each case are: Control Group ($M = 8.00$) and Experimental Group ($M = 23.00$). The relative learning gain between the Control Group and the Experimental Group are statistically significant different, p-value less than 0.05 ($P = .000001$). The effect size (r) is calculated using the (Equation 2) and found that $r = 0.8$ indicating a large effect size using (Cohen, 1988) criteria. It can be concluded that the average improvement in test scores in the Experimental Group are statistically significant comparing to the Control Group.

Table 7.11 and 7.12 show the results of Wilcoxon Signed Ranks test conducted to test the students' performance between the three learning styles (V: Visual learning style, A: Auditory learning style, K: kinaesthetic learning style).

Table 7.11: the differences between the mean between the three learning style (V: visual, A: auditory, K: kinaesthetic))

Descriptive Statistics								
	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
Post-score_K	5	10.00	.000	10	10	10.00	10.00	10.00
Post-score_V	5	9.6000	.54772	9.00	10.00	9.0000	10.0000	10.0000
Post-score_A	5	9.8000	.44721	9.00	10.00	9.5000	10.0000	10.0000
Pre-score_K	5	4.60	1.517	3	7	3.50	4.00	6.00
Pre-score_V	5	3.6000	2.07364	1.00	6.00	1.5000	4.0000	5.5000
Pre-score_A	5	4.4000	1.14018	3.00	6.00	3.5000	4.0000	5.5000

Table 7.12: The P-Values between the means from table 7.10.

Test Statistics ^a			
	Post_scoreK - Pre_scoreK	Post_scoreV - Pre_scoreV	Post_scoreA - Pre_scoreA
Z	-2.032 ^b	-2.032 ^b	-2.041 ^b
Asymp. Sig. (2-tailed)	.042	.042	.041
a. Wilcoxon Signed Ranks Test			
b. Based on negative ranks.			

It can be seen from table 7.11 that there are differences between the means (the pre-test and post-test scores) for each learning style. The mean of the post-test scores are higher for all the learning style (V, A, and K) than the pre-test scores which suggests that the students performed well in the post test after adapting to their learning style. Table 7.12 illustrates the P-Value for each learning style, which highlighted that the difference between the pre-test and post-test scores is statistically significant (less than P-Value .05) for all three leaning styles (Visual: P = .042, Auditory: P=.041, kinaesthetic: P=.042).

7.5.3.2 Adapting VAK model to autistics learning style (Questionnaire) – Results and Discussion

Table 7.13, Table 7.14 and Table 7.15 demonstrate the results of the third statement in the questionnaire to test the participant's perception of remembering what LANA-

II has taught them between Control Group (without using VAK) and Experimental Group (with using VAK).

Table 7.13: Crosstab *with and without VAK - Questiona3: I think I remember what LANA-II has taught me in the tutorial.

			VAK		Total
			With	Without	
I think I remember what LANA has taught me in the tutorial	Happy	Count	14	8	22
		% within VAK	93.3%	53.3%	73.3%
	Normal	Count	1	7	8
		% within VAK	6.7%	46.7%	26.7%
	Sad	Count	0	0	0
		% within VAK	0%	0%	0%
Total		Count	15	15	30
		% within VAK	100.0%	100.0%	100.0%

Table 7.14: Chi-Square Test **with and without VAK - Questiona3: I think I remember what LANA has taught me in the tutorial.

Chi-Square Tests with and without VAK					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	6.136 ^a	1	.013		
Continuity Correction ^b	4.261	1	.039		
Likelihood Ratio	6.719	1	.010		
Fisher's Exact Test				.035	.018
N of Valid Cases	30				
a. 2 cells (50.0%) have expected count less than 5. The minimum expected count is 4.00.					
b. Computed only for a 2x2 table					

Table 7.15: The effect size table for question 3..

Symmetric Measures			
		Value	Approx. Sig.
Nominal by Nominal	Phi	.452	.013
	Cramer's V	.452	.013
N of Valid Cases		30	

Table 7.13 illustrates that the majority of students in Experimental Group answered they remembered the content of the tutorial (93.3%). (6.7%) of the students have a neutral feeling (i.e. selected Normal face on the questionnaire); whereas students in

Control Group were remembering the tutorial with a percentage of (53.3%) and (46.7%) stated they have a neutral feeling when rating this question. A Chi-Square Test in table 7.14 shows that there is statistically significant relationship between (adapting, not adapting) the learning style and answering this question (P- value less than .05) $P=.013$. Table 7.15 shows the effect size in this test, $\Phi = 0.45 \approx 0.5$ indicating a large effect size using (Cohen, 1988) criteria.

Figure 7.2 shows that the students in Experimental Group (adapting to VAK learning style) were happier in remembering the tutorial than the students in Control Group (without adapting to VAK learning style). This indicates that the students in Experimental Group believed that using their leaning style helped them to remember the tutorial.

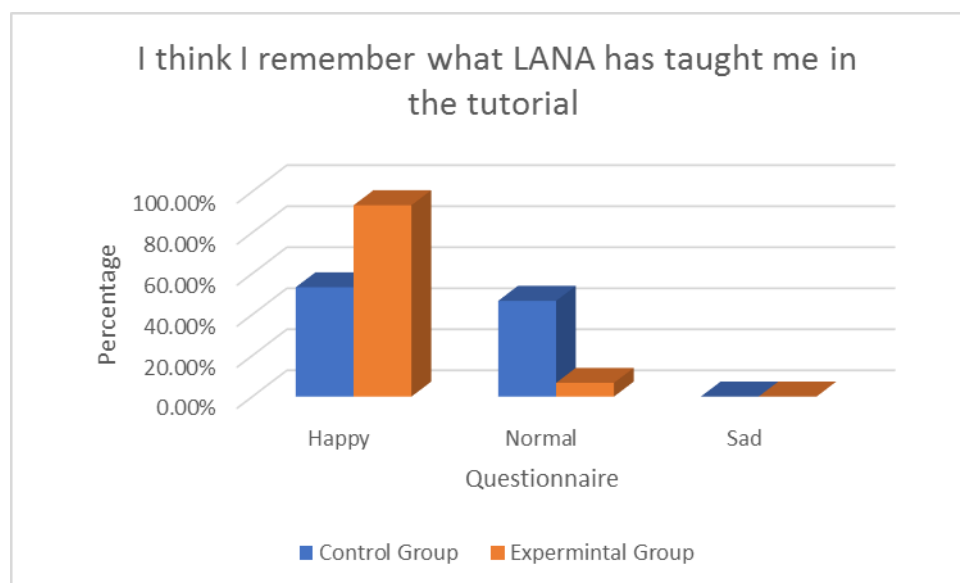


Figure 7.2: Question 3 between Control Group and Experimental Group

Table 7.16 shows the result of the fourth question in the questionnaire to test the participant's perception of LANA-II adapting to their learning style in the tutorial.

Table 7.16: Crosstab Questiona4: I think I can learn better with adapting to my learning style (Visual – Auditory -Kinaesthetic).

			VAK	Total
			With	
I think I can learn better with using my learning style (Visual – Auditory - Kinaesthetic)	Happy	Count	14	14
		% within VAK	93.3%	93.3%
	Normal	Count	1	1
		% within VAK	6.7%	6.7%
	Sad	Count	0	0
		% within VAK	0%	0%
Total		Count	15	15
		% within VAK	100.0%	100.0%

The results from Table 7.16 and Figure 7.3 highlighted that the majority of the students are happy with adapting to their learning style with percentage (93.3%), whereas (6.7%) of the student have a neutral feeling when rating this question.



Figure 7.3: Question 4 analysis bar chart

Table 7.17, Table 7.18 and Table 7.19 illustrate the result of the sixth question in the questionnaire to test the participant's perception of learning from LANA-II more than leaning from the teacher.

Table 7.17: Crosstab *with and without VAK - Question 6: I think I can learn from LANA-II better than learning from my teacher.

			VAK		Total
			With	Without	
I think I can learn from LANA better than my teacher	Happy	Count	14	9	23
		% within VAK	93.3%	60.0%	76.7%
	Normal	Count	1	6	7
		% within VAK	6.7%	40.0%	23.3%
	Sad	Count	0	0	0
		% within VAK	0%	0%	0%
Total		Count	15	15	30
		% within VAK	100.0%	100.0%	100.0%

Table 7.18: Chi-Square Test *with and without VAK – Question 6: I think I can learn from LANA-II better than learning from my teacher.

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	4.658 ^a	1	.031		
Continuity Correction ^b	2.981	1	.084		
Likelihood Ratio	5.058	1	.025		
Fisher's Exact Test				.080	.040
N of Valid Cases	30				

a. 2 cells (50.0%) have expected count less than 5. The minimum expected count is 3.50.

b. Computed only for a 2x2 table

Table 7.19: The effect size table for Question 6.

Symmetric Measures			
		Value	Approx. Sig.
Nominal by Nominal	Phi	.394	.031
	Cramer's V	.394	.031
N of Valid Cases		30	

The results show that the majority of students in the Experimental Group were preferring to learn from LANA-II than learning from the teacher with percentage of (93.3%), and (6.7%) of the students have a neutral feeling. The students in the Control Group were preferring to learn from LANA-II than learning from the teacher with percentage of (60%), and (40%) stated they have a neutral feeling when rating this question. A Chi-Square Test in table 7.18 shows that there is statistically significant relationship between (adapting, not adapting) the learning style and answering this

question (P- value less than .05) $P=.031$. Table 7.19 shows the effect size in this test, $\Phi = 0.39 \approx 0.4$ indicating a nearly large effect size using (Cohen, 1988) criteria.

Figure 7.4 shows that the students in the Experimental Group (adapting to VAK learning style) are happier in learning from LANA-II than the students in the Control Group (without adapting to their VAK learning style). This indicates that student's leaning style has a strong effect in the student's learning especially when material is presented to them in accordance to their learning style.

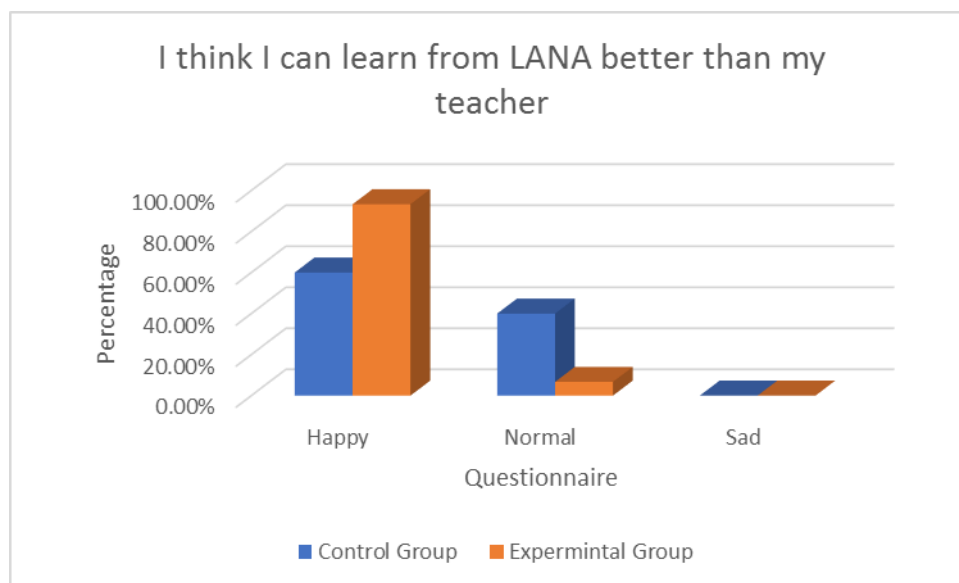


Figure 7.4: Question 6 between Control Group and Experimental Group

7.5.4 Experiment 1 Conclusion

As stated in section 7.5.2, the experimental results evaluated four metrics related to Hypothesis A:

1. Pre-test and Post-test with and without adapting to the VAK learning style.
2. The participant's perception of remembering the tutorial with and without adapting to VAK learning style.
3. The participant's perception of the tutorial adaption in accordance with their learning style.
4. The participant's perception of learning the tutorial from LANA-II more effectively than from the teacher.

It can be summarized from the results that LANA-II can be successfully adapted to the learning styles of autistic students. As illustrated in section (7.5.3.1), there is a significant difference in results between (the pre-test and post-test scores) between

adapting to the VAK learning style and without adapting to VAK ($p\text{-value}=0.000001$). This provided evidence that the learning gain of the Experimental Group is improved over the Control Group.

The results of the participant's perception of remembering the tutorial were measured through the user questionnaire. The results showed that the students that were adapted to their VAK leaning style were able to remember the tutorial (93.3%) more than the students that are not adapted to their VAK learning style (53.3%) (Section 7.5.3.2). The participant's perception of adapting their learning style was indicated using the user questionnaire. The results found that the majority of the students were happy with adapting to their learning style with the tutorial (93.3%) (Section 7.5.3.2). The participant's perception of learning using LANA-II rather than learning from the teacher was measured using the user questionnaire. It can be seen from the results (section 7.5.3.2) that the students that used a VAK leaning style had greater preference to learn from LANA-II (93.3%) than the students that are not adapted to their VAK learning style (60%). Therefore, the main hypothesis (LANA-II using VAK model can adapt to the Autistics learning style) can be accepted.

7.6 Experiment 2 – LANA-II engine effectiveness

7.6.1 Participants

The total size of the sample was 24 participants with age group (10-12) years old and their first language was Arabic. The participants recruited for the evaluation are from the UK and none of them had a previous experience using LANA-I or LANA-II. All of the participant's parents received an information sheet about the project and its aims, and a consent form to get their permission before conducting the experiment (see Appendix D and Appendix E for the information sheet and consent form). The participants were divided into two equal sized groups. The first group was the Control Group (12 participants) used the LANA-II without the adaptation to VAK learning style model. The second group was the Experimental Group (12 participants) used LANA-II with the adaptation to VAK learning style model.

7.6.2 Experiment 2 Methodology

This experiment was conducted to test the hypothesis HB_1 based on the metrics outlined in section 7.2.3. The data for this experiment was gathered from the LANA-

II log file and the user feedback questionnaire. The subjective and objective metrics (Table 7.20) were used to answer questions related to Hypothesis B (The enhancements made to LANA-II architecture improve the overall effectiveness of LANA-II engine):

Question 1: Is the LANA-II engine more effective than the (LANA-I) engine?

Question 2: Do the improvements made to the LANA-II engine result in improved perception over (LANA-I) by the user?

The data gathered from the log file will allow assessment of the performance of the LANA-II components and the algorithms deployed in the architecture (Chapter 6). This data will be compared with the first set of data (LANA-I - Chapter 5).

Table 7.20: Experiment 2 metrics

Metrics to be Evaluated	Method of Evaluation	Hypothesis
QQ1: LANA Tutor is enjoyable to use	Questionnaire	Hypothesis B
QQ2: LANA has a right amount of information in one Tutorial.	Questionnaire	Hypothesis B
QQ5: I like the design and colour of the interface.	Questionnaire	Hypothesis B
QQ7: I would like to use LANA again with different topics.	Questionnaire	Hypothesis B
QQ8: The conversation was natural.	Questionnaire	Hypothesis B
QQ6: I think I would like to learn using LANA better than the teacher in class	Questionnaire	Hypothesis B
Completion time	Log file	Hypothesis B
Number of unrecognised Utterance	Log file	Hypothesis B
Number of correct and incorrect response	Log file	Hypothesis B
Similarity Strength	Log file	Hypothesis B
The rule fired	Log file	Hypothesis B

7.6.3 Experiment 2 Results and Discussion

7.6.3.1 Log file analysis

The data from the LANA-II log file was collated, processed and analysed. Table 7.21 summarises the results of data gathered through the log file during the LANA-II and LANA-I evaluation (see Chapter 5). These results show that LANA-II performed better during the end user evaluation. Further analysis is done for these results in the following sections to determine whether these results are significantly different from the results of the LANA-I.

Table 7.21: Log file analysis from the second and first prototype

CONVERSATION ANALYSIS		
METRIC	LANA-II	LANA-I
Total number of conversations	24	24
Total number of utterances in all conversations	310	313
Average number of words per user utterance	3.1	3.1
Average number of utterances per conversation	13	13
Percentage of correct response	96.4%	89.8%
Percentage of incorrect response	3.6%	10.2%
Percentage of unrecognized utterance because of the knowledge gap	5.4%	38%
Percentage of utterances using Similarity Strength	42.7%	57.18%
Percentage of utterances have corrected response using STS algorithm	42.7%	19%
Number of the rule fired	280	170
Average conversation duration (mins)	12	11.5

7.6.3.2 Objective data analysis from old and new data sets

This section illustrates the objective data analysis to find out the differences in the data in order to test hypothesis **HB₁** (The enhancement made to the LANA-II architecture improve the overall effectiveness of LANA-II engine). Mann-Whitney U Tests are used to test the significance of the results.

7.6.3.3 Number of unrecognised utterances

Table 7.22 shows the results of the Mann-Whitney U test to determine if there was any difference in number of unrecognised utterances between LANA-I and LANA-II.

Table 7.22: Mann-Whitney U test for unrecognised utterances

Ranks				
		N	Mean Rank	Sum of Ranks
Number of unrecognised utterances	LANA-I	24	36.50	876.00
	LANA-II	24	12.50	300.00
	Total	48		
Test Statistics^a				
	Number of unrecognised utterances			
Mann-Whitney U	.000			
Wilcoxon W	300.000			
Z	-6.371			
Asymp. Sig. (2-tailed)	.000			

Table 7.22 shows that there is a statistically significant difference in the number of unrecognised utterances between LANA-I and LANA-II, p-values less than 0.05 (reported by the SPSS as $p = .000$). When comparing the mean ranks, of the two systems LANA-I was ranked higher, which means LANA-I having more unrecognised

utterances. The effect size (r) is calculated using the (Equation 2) and found that $r = 0.9$ indicating a very large effect size using (Cohen, 1988) criteria.

7.6.3.4 Number of correct responses

Table 7.23 illustrates the results of the Mann-Whitney U test conducted in order to measure whether there was a statistically significant difference between the numbers of correct responses between LANA-I and LANA-II.

Table 7.23: Mann-Whitney U test for correct responses

		N	Mean Rank	Sum of Ranks
Number of correct response	LANA-I	24	17.67	424.00
	LANA-II	24	31.33	752.00
	Total	48		
Test Statistics^a				
				Number of correct response
Mann-Whitney U				124.000
Wilcoxon W				424.000
Z				-3.911
Asymp. Sig. (2-tailed)				0.000092

From this result, it can be concluded that the number of correct responses between the LANA-I and LANA-II was statistically significant ($p = 0.000092$). In comparing the mean rank LANA-II ranks higher, which indicates that LANA-II having more correct responses than LANA-I. The effect size (r) is calculated using the (Eq1) and found that $r = 0.5$ indicating a large effect size using (Cohen, 1988) criteria.

7.6.3.5 Number of incorrect responses

Table 7.24 outlines the results of a Mann-Whitney U test conducted to test if there was a statistically significant difference between the numbers of incorrect responses between LANA-I and LANA-II.

Table 7.24: Mann-Whitney U test for incorreceted responses

		N	Mean Rank	Sum of Ranks
Number of incorrect response	LANA-I	24	30.56	733.50
	LANA-II	24	18.44	442.50
	Total	48		
Test Statistics^a				
				Number of incorrect response
Mann-Whitney U				142.500
Wilcoxon W				442.500
Z				-3.237
Asymp. Sig. (2-tailed)				.001

The results indicated that there are statistically significant differences between the numbers of the incorrect response between LANA-I and LANA-II ($p = 0.001$). When

comparing the mean ranks it can be seen that LANA-I ranks higher than LANA-II, meaning that LANA-I having more incorrect responses than LANA-II. The effect size (r) is calculated using the (Equation 2) and found that $r = 0.4$ indicating a nearly large to large effect size using (Cohen, 1988) criteria.

7.6.3.6 Number of utterances have corrected response using STS algorithm

Table 7.25 illustrates the results of a Mann-Whitney U test conducted to test if there was a statistically significant difference between the numbers of utterances have correct response using STS algorithm between LANA-I and LANA-II.

Table 7.25: Mann-Whitney U test for correct responses using STS algorithm

Ranks				
	Prototype	N	Mean Rank	Sum of Ranks
Number of utterances have correct response using STS algorithm	LANA-I	24	15.25	366.00
	LANA-II	24	33.75	810.00
	Total	48		
Test Statistics ^a				
		Number of utterances have corrected response using STS algorithm		
Mann-Whitney U		66.000		
Wilcoxon W		366.000		
Z		-5.194		
Asymp. Sig. (2-tailed)		.000		
a. Grouping Variable: Prototype				

From this result, it can be deduced that there is a statistically significant difference between LANA-I and LANA-II (reported by the SPSS as $p = .000$). When comparing the mean ranks it can be seen that LANA-II ranks higher than LANA-I, meaning that the correct responses using the new STS algorithm in LANA-II was significantly improved compared to the correct responses using STS algorithm of LANA-I. The effect size (r) is calculated using the (Equation 2) and found that $r = 0.7$ indicating a very large effect size using (Cohen, 1988) criteria.

7.6.3.7 Number of rules fired

Table 7.26 illustrated the results of a Mann-Whitney U test conducted to test if there was a statistically significant difference between the numbers of rules fired between LANA-I and LANA-II for the same tutorial (solar system tutorial, chapter 4).

Table 7.26: Mann-Whitney U test for the number of the rules fired

Table A10: Mann-Whitney U test for the number of the rules fired				
		N	Mean Rank	Sum of Ranks
Number of rule fired	LANA-I	24	12.50	300.00
	LANA-II	24	36.50	876.00
	Total	48		
Test Statistics ^a				
		Rule_fired		
Mann-Whitney U		.000		
Wilcoxon W		300.000		
Z		-6.395		
Asymp. Sig. (2-tailed)		.000		
a. Grouping Variable: Prototype				

The results indicated that there are statistically significant differences between the numbers of the rules fired between LANA-I and LANA-II (reported by the SPSS as $p = .000$). When comparing the mean ranks it can be seen that LANA-II ranks higher than LANA-I. This means that for the same tutorial, LANA-II had more fired rules than LANA-I. The effect size (r) is calculated using the (Equation 2) and found that $r = 0.9$ indicating a very large effect size using (Cohen, 1988) criteria.

7.6.3.8 Number of utterances using Similarity Strength

Table 7.27 shows the results of a Mann-Whitney U test to test if there was a statistically significant difference between the numbers of utterances using similarity strength between LANA-I and LANA-II. This result is for all utterances that used STS algorithm (which have correct and incorrect responses).

Table 7.27: Mann-Whitney U test for the numbers of utterances using similarity strength.

Ranks				
		N	Mean Rank	Sum of Ranks
Number of utterances using Similarity Strength	LANA-I	24	27.08	650.00
	LANA-II	24	21.92	526.00
	Total	48		
Test Statistics ^a				
	Number of utterances using Similarity Strength			
Mann-Whitney U	226.000			
Wilcoxon W	526.000			
Z	-1.369			
Asymp. Sig. (2-tailed)	.171			
a. Grouping Variable: Prototype				

The results illustrate that there is not a statistically significant ($p = .171$) difference between the number of user utterances which using similarity strength between the LANA-I and LANA-II.

7.6.3.9 Time taken to complete the tutorial

Table 7.28 illustrates the results of the Mann-Whitney U test conducted in order to measure whether there was a statistically significant difference between the times taken to complete the tutorial between LANA-I and LANA-II.

Table 7.28: Mann-Whitney U test for the average of the conversation duration.

Ranks				
		N	Mean Rank	Sum of Ranks
Average conversation duration (mins)	LANA-I	24	23.00	552.00
	LANA-II	24	26.00	624.00
	Total	48		
Test Statistics ^a				
	Average conversation duration (mins)			
Mann-Whitney U	252.000			
Wilcoxon W	552.000			
Z	-.864			
Asymp. Sig. (2-tailed)	.388			
a. Grouping Variable: Prototype				

From this result, it can be concluded that the time taken to reach the conversation goal between LANA-I and LANA-II was not statistically significant ($p = 0.388$).

This concludes the results analysis of the log file data. The following section explores and analyses the questionnaire data that was gathered in order to gauge participants perceptions related to the subjective metrics.

7.6.3.10 Analysis of questionnaire data

Table 7.29 summarises the findings of the questionnaire from the evaluation of LANA-I and Table 7.30 outlines a summary of the results of the questionnaire from the evaluation of the LANA-II. Table 7.31 and Table 7.32 shows the means values of the questionnaire one and two.

Table 7.29: Frequency analysis LANA-I questionnaire data results

QUESTIONNAIRE RESULTS	Happy	Normal	Sad
I think LANA Tutor is enjoyable to use	100%	0	0
I think LANA had a right amount of information in one Tutorial.	91.6%	8.4%	0
I like the design and colour of LANA.	91.6%	8.4%	0
I think I remember what LANA has taught me in the tutorial.	91.7%	8.3%	0
I think I can learn better with using my learning style (Visual – Auditory -Kinaesthetic).	83.3%	16.7%	0
I think I would like to use LANA again with other lessons.	91.6%	8.4%	0
I think talking with LANA-Is like as talking with my friend.	45.8%	50%	4.2%
I think I would like to learn using LANA better than the teacher in class	79.16%	20.84%	0

Table 7.30: Frequency analysis LANA-II questionnaire data results

QUESTIONNAIRE RESULTS	Happy	Normal	Sad
I think LANA Tutor is enjoyable to use	100%	0	0
I think LANA had a right amount of information in one Tutorial.	91.6%	8.4%	0
I like the design and colour of LANA.	100%	0	0
I think I remember what LANA has taught me in the tutorial.	91.7%	8.3%	0
I think I can learn better with using my learning style (Visual – Auditory -Kinaesthetic).	91.6%	8.4%	0
I think I would like to use LANA again with other lessons.	91.6%	8.4%	0
I think talking with LANA-Is like as talking with my friend.	91.6%	8.4%	0
I think I would like to learn using LANA better than the teacher in class	91.6%	8.4%	0

Table 7.31: Mean values from evaluation questionnaire LANA-I

Report							
Mean							
Enjoyable	Satisfaction with the information.	Remembering and understanding	Satisfaction with learning style (Visual – Auditory - Kinaesthetic)	Design and colour	Learning from LANA-Is better than the teacher.	Using LANA again with other lessons.	Naturalness
3.0000	2.9583	2.9583	2.9167	2.9583	2.7917	2.9583	2.4583

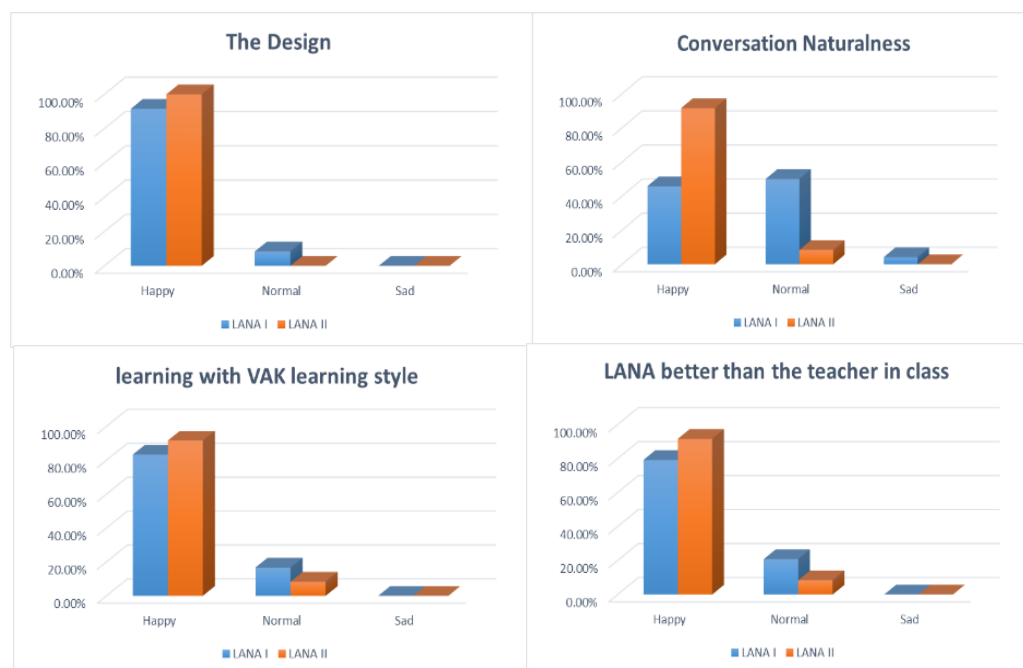
Table 7.32: Mean values from evaluation questionnaire LANA-II

Mean							
Enjoyable	Satisfaction with the information.	Remembering and understanding	Satisfaction with learning style (Visual – Auditory - Kinaesthetic)	Design and colour	Learning from LANA-Is better than the teacher.	Using LANA again with other lessons.	Naturalness
3.0000	2.9583	2.9583	2.9583	3.0000	2.9167	2.9583	2.9583

Table 7.33: Mann Whitney test between LANA-I and LANA-II questionnaire data.

Test Statistics ^a								
	Enjoyable	Satisfaction with the information .	Remembering and understanding	Satisfaction with learning style	Design and colour	Learning from LANA-Is better than the teacher.	Using LANA again with other lessons.	Naturalness
Mann-Whitney U	288.000	288.000	288.000	276.000	276.000	240.000	288.000	144.000
Wilcoxon W	588.000	588.000	588.000	576.000	576.000	540.000	588.000	444.000
Z	.000	.000	.000	-.590	-1.000	-1.727	.000	-3.771
Asymp. Sig. (2-tailed)	1.000	1.000	1.000	.555	.317	.084	1.000	.000163
a. Grouping Variable: Prototype								

Table 7.33 illustrates the results of the Mann Whitney test conducted on the matching questions from the questionnaires administered in the LANA-I and LANA-II evaluations. The results show that there is one significant differences in question 8 (the naturalness) $p\text{-value} = 0.000163$. The effect size (r) is calculated using the (Equation 2) and found that $r = 0.5$ indicating a large effect size using (Cohen, 1988) criteria. When comparing the mean values from (LANA-I) questionnaire and LANA-II questionnaire, it is found that four questions (Q1,Q2,Q3,Q7) from both questionnaires have the same mean values whereas the mean values of four questions (Q4,Q5,Q6,Q8) in questionnaire two are higher than (LANA-I) questionnaire which indicated an improvement in these questions rating (see Figure 7.5).

**Figure 7.5: Bar Chart questionnaire results between LANA-I and LANA-II**

In the next section, the collected data from LANA-II will be compared between two groups participants (Autistic and Neurotypical children) in order to highlight any differences that may be present between these data. This will provide further insights in to whether or not the autistic participants and neurotypical participants demonstrated any significant differences during their interaction with LANA-II during the evaluation.

7.6.4 Comparative descriptive analysis of data from autistic participants and neurotypical participants

This section presents further descriptive analysis of the collected data through the log file. This data was collected from all autistic students (30 participants) and all neurotypical students (24 participants). Both participants used LANA-II in both tutorials (Solar system, Electrical circuit and the magnetism). The analysis of results presented in this part are intended to highlight differences of the results between autistic participants and neurotypical participants. The Man-Whitney U test was used in order to highlight any statistically significant differences in the data. The findings are as follows:

7.6.4.1 Duration of conversation whilst engaged with tutorial

Table 7.34 illustrates the results of a Mann-Whitney U test conducted to test whether there was a difference in the time taken to complete the tutorial between the two groups participants.

Table 7.34: Mann-Whitney U test for duration of conversation between the two groups participants

participants				
Ranks				
	Participants	N	Mean Rank	Sum of Ranks
Average conversation duration (mins)	NT	24	12.50	300.00
	Autistic	30	39.50	1185.00
	Total	54		
Test Statistics ^a				
	Average conversation duration (mins)			
Mann-Whitney U	.000			
Wilcoxon W	300.000			
Z	-6.866			
Asymp. Sig. (2-tailed)	.000			
a. Grouping Variable: Participants				

The results show that there is a statistically significant difference in the time taken to complete the tutorial between the two groups participants (reported by the SPSS as p

= .000). When the mean ranks are compared, the results show that autistic participants rank higher than the neurotypical participants, indicating that the autistic participants spent time more than the neurotypical participants to complete the tutorial. The effect size (r) is calculated using the (Equation 2) and found that $r = 0.9$ indicating a very large effect size using (Cohen, 1988) criteria.

7.6.4.2 Number of unrecognised utterances

Table 7.35 shows the results of the Mann-Whitney U test to determine if there was any difference in number of unrecognised utterances between the two groups participants.

Table 7.35: Mann-Whitney U test for the number of unrecognised utterances between the two groups participants

Ranks				
	Participants	N	Mean Rank	Sum of Ranks
Number of unrecognised utterance	NT	24	31.63	759.00
	Autistic	30	24.20	726.00
	Total	54		
Test Statistics ^a				
	Number of unrecognized utterance			
Mann-Whitney U	261.000			
Wilcoxon W	726.000			
Z	-2.002			
Asymp. Sig. (2-tailed)	.045			
a. Grouping Variable: Participants				

The results show that there is a statistically significant difference in the number of unrecognized utterances between the two groups participants ($p = .045$). When the mean ranks are compared the results show that neurotypical participants ranks higher than the autistic participants, indicating that the number of unrecognised utterances from neurotypical participants was more than the autistic participants. The effect size (r) is calculated using the (Equation 2) and found that $r = 0.3$ indicating a medium effect size using (Cohen, 1988) criteria.

7.6.4.3 Number of correct responses

Table 7.36 illustrate the results of the Mann-Whitney U test conducted in order to measure whether there was a statistically significant difference between the numbers of correct responses between the two groups participants.

Table 7.36: Mann-Whitney U test for the number of correct utterances between the two groups participants

Participants				
Ranks				
	Participants	N	Mean Rank	Sum of Ranks
Number of correct response	NT	24	22.81	547.50
	Autistic	30	31.25	937.50
	Total	54		
Test Statistics ^a				
	Number of correct response			
Mann-Whitney U	247.500			
Wilcoxon W	547.500			
Z	-2.560			
Asymp. Sig. (2-tailed)	.010			
a. Grouping Variable: Participants				

From this result, it can be concluded that the number of correct responses between the two groups participants was statistically significant ($p = .010$). In comparing the mean ranks the autistic participants rank higher, which indicates that the autistic participants having more correct responses than the neurotypical participants. The effect size (r) is calculated using the (Equation 2) and found that $r = 0.3$ indicating a medium effect size using (Cohen, 1988) criteria.

7.6.4.4 Number of incorrect responses

Table 7.37 outlines the results of a Mann-Whitney U test conducted to test if there was a statistically significant difference between the numbers of incorrect responses between the two groups participants.

Table 7.37: Mann-Whitney U test for the number of incorrect utterances between the two groups participants

Ranks				
	Participants	N	Mean Rank	Sum of Ranks
Number of incorrect response	NT	24	32.38	777.00
	Autistic	30	23.60	708.00
	Total	54		
Test Statistics ^a				
	Number of incorrect response			
Mann-Whitney U	243.000			
Wilcoxon W	708.000			
Z	-2.625			
Asymp. Sig. (2-tailed)	.009			
a. Grouping Variable: Participants				

The results indicated that there were statistically significant differences between the numbers of the incorrect response between the autistic participants and the neurotypical participants ($p = 0.009$). When comparing the mean ranks it can be seen that the neurotypical participants ranks higher than autistic participants, meaning the

neurotypical participants having more incorrect responses than the autistic participants. The effect size (r) is calculated using the (Equation 2) and found that $r = 0.4$ indicating a nearly large effect size using (Cohen, 1988) criteria.

7.6.4.5 Number of utterances using Similarity Strength

Table 7.38 shows the results of a Mann-Whitney U test to test if there was a statistically significant difference between the numbers of utterances using similarity strength between the two groups participants.

Table 7.38: Mann-Whitney U test for the numbers of utterances using similarity strength.

Ranks				
	Participants	N	Mean Rank	Sum of Ranks
Number of utterances using Similarity Strength	NT	24	39.54	949.00
	Autistic	30	17.87	536.00
	Total	54		
Test Statistics ^a				
	Number of utterances using Similarity Strength			
Mann-Whitney U	71.000			
Wilcoxon W	536.000			
Z	-5.327			
Asymp. Sig. (2-tailed)	.000			
a. Grouping Variable: Participants				

From this result, it can be summarized that there is a statistically significant difference between the two groups participants (reported by the SPSS as $p = .000$). When comparing the mean ranks it can be seen that the neurotypical participants ranks higher than autistic participants, meaning that the number of utterances used the similarity strength in neurotypical participants was more than the number in autistic participants, which means most of the autistic student utterances were matched the scripted patterns. The effect size (r) is calculated using the (Equation 2) and found that $r = 0.7$ indicating a very large effect size using (Cohen, 1988) criteria.

7.7 Case Study

This case study was conducted to test the hypothesis HC_1 based on the metrics outlined in section 7.2.3. The data for this experiment was gathered from interviews and observations (section 7.7.3.1).

7.7.1 Participants

There are no rules for sample size in a qualitative inquiry (Patton, 2005) the sample size is dependent upon the purpose of the study, access to participants, and time and

resources (Patton, 2005). In this research study, three participants who met the following criteria were interviewed and observed:

- Participant was within the age group (10-12) years old.
- He/she had previously received a community clinical diagnosis of high functioning autism spectrum disorder or Asperger's, no language problem and intellectual impairments. This was determined by the classroom teacher.
- His/her first language was Arabic.
- The participant was studying in mainstream primary schools in Saudi Arabia.
- They had no previous experience using LANA CITS.

Parents received an information sheet about the project and its aims, and a consent form to get their permission before conducting the experiment (see Appendix D and Appendix E for the information sheet and consent form). To ensure data privacy, the names of participants have been changed in this study.

In this study, three autistic students are involved and all of them are female (Mariam, Fatimah, and Dana). Two of them (Mariam and Fatimah) are in (Grade 5), and their age is 11 years old, the third students (Dana) is in (Grade 4) and her age is 10 years old. The three participants are studying in mainstream primary school, providing care for children with ASD in Saudi Arabia. All the three participants have a first language of Arabic and none of them had a previous experience using LANA CITS. Each child had previously received a community clinical diagnosis of Asperger's, who they have no problem with language and intellectual impairments.

7.7.2 Ethical Considerations

Because the participants were children, a consent form and information sheet were sent to the participant's parents (see Appendix E). The information sheet outlined the details and purpose of the study, and the benefits for participating in the study. They also informed that the information gathered from the participant is stored securely.

7.7.3 Data Collection Methods

There are several data collections methods in qualitative research. (Marshall and Rossman, 2006) identified three primary methods that qualitative researchers use for gathering information. Primary methods involve gathering information from the

source directly. These three methods are: 1) participating in the setting, 2) observing directly, and 3) interviewing in depth. In this study, two primary methods are used: Observation and Interviewing. “Observation involves the systematic noting and recording of events, behaviors, and artefacts in the social setting,” (Marshall and Rossman, 2006). (Kvale and Brinkmann, 2009) say that an interview is a conversation with structure and a purpose. There are secondary methods that may also be used to gather information. This involves gathering information from someone or something else outside of the original source by analysing documents such as: videos, photography, interaction analysis, and questionnaires (Marshall and Rossman, 2006).

7.7.3.1 Interviewing

Interviewing is a universal method in qualitative research (Byrne, 2004; Patton, 2005) Interviews have been divided into three categories: structured, unstructured, and semi-structured. In structured interviews, all participants will have the same questions using the same wording and in the same order (Corbetta, 2003). However, while a structured interview ensures that comparable responses are gathered from each participant, it leaves no room for elaboration (Berg, 2004).

The second category is an unstructured interview, which contains open questions and subsequent questions depending on the participant’s responses (Holloway and Galvin, 2016). The drawbacks of this type of interview are the difficulty of processing data and the effort required. It is difficult to find similar statements from different participants and hence the links are often difficult to make.

The most common type of interviews used in qualitative research are the third category, which is semi-structured interviews (Holloway and Galvin, 2016). It involves structured questions and the researcher is free vary the order and wording of the questions (Power et al., 2010), depending on the direction of the interview, and to ask additional questions (Corbetta, 2003).

In this study, two interviews were conducted. The first interview was with the autistic student participant and the second interview was with the teacher. The first interview, which was with the autistic student, contains nine questions and the interviewer was free to ask follow-up questions. Table 7.39 shows the nine questions that are used in the interview.

Table 7.39: The Interview Sheet for the autistic student (translated version).

1. Do you think LANA Tutor is enjoyable to use? • Which part do you like? • Why do you like this part?
2. Do you think LANA had a right amount of information in one Tutorial?
3. Do you remember what LANA has taught you in the tutorial? • Tell me in one or two things you was learned from LANA
4. Did you like using your learning style (Visual – Auditory -Kinaesthetic) in LANA? • Would you like to use your learning style with all subjects and tutorials?
5. Do you like the way LANA looked? • What do you think of the colour of LANA?. • Any suggestion to add in the design or the character.
6. Do you think you can learn from LANA better than your teacher?
7. Would you like to use LANA again with other lessons?
8. Is talking with LANA like talking with your friend?
9. Would you like to use LANA-In the classroom?

According to (Berg, 2004), the researcher is able to adjust the wording of the questions, and add or delete probes, as necessary. With this structure, the interviewer guided the interview, to obtain a good description from the participant's perspective. Moreover, this structure allowed the researcher the flexibility to dive deeper when appropriate. (Patton, 2005) asserted that the semi-structured process lets the researcher to explore and ask follow-up questions if necessary to clarify the particular subject. This was an extremely beneficial process in the field of qualitative interviewing. The use of semi-structured interviewing was important to this study because the participants are children and the level of comprehension may vary. The interview was interactive and the participants were able to ask questions. In addition, notes were taken during the interview to capture the core of the participant's comments and to note any non-verbal behaviour and facial expressions.

The second interview was conducted with the teacher. The aim of this interview is to measure the student's interaction and behaviour from the teacher's point of view. The interview contains five questions as shown in Table 7.40.

Table 7.40: The Interview Sheet with the teacher (translated version).

1. Is there any behavioural issue with the student?
2. What is the student's education level in the school?
3. How was the student's interaction during the tutorial?
4. Did the student interact with the LANA's tutorial in the same way as the class tutorial?
5. Do you have any suggestions to add to LANA?

7.7.3.2 Participant Observation

In addition to semi-structured interviews, the participant was been observed during the interview by the teacher and the researcher. Observations allow the interviewer to observe the context in which the participants interact (Patton, 2005). Additionally, observations allow the researcher to notice the participant's behaviors, facial expressions, and the interactions.

7.7.4 Data Analysis

Following the interviews, the data was analysed using the following methodology (Creswell et al., 2007):

1. Data management: This involves organizing and gathering all the data. In this study, the interview was transferred into a written format.
2. Reading and noting: This involves reading through all the transcriptions and making notes in the margins regarding the interviewee's comments.
3. Describing: this involves describing the participant's experience with the study.
4. Classifying: The researcher then makes a list of significant statements, quotes that emerged from the data and then grouped them together.
5. Interpreting: the researcher develops textural and structural descriptions of the data. Textural descriptions are the written descriptions of what the participants experienced (Creswell et al., 2007). After the textural descriptions are written, the researcher outlined the contextual influence, and this is the structural description.
6. Representing and visualizing: After both the textual and structural descriptions are written, the researcher writes the findings of the study. In addition, tables were created to help illustrate the data (Creswell et al., 2007).

7.7.5 Participant's interaction

The study started by detecting the participant's learning style using the VAK questionnaire, which was answered by both the student and the teacher. The registration interface started to register the participant with her learning style in order to adapt the tutorial according to her learning style. Then, a pre-test interface started and the participant was asked to complete the pre-test by choosing the right answer using the touch screen. When the test was completed, the tutorial started and the participant interacted with the tutorial and answered the questions. During the tutorial, materials were used according to the learning style (Visual: video and images, Auditory: sounds, Kinaesthetic: models). After the participant completed the tutorial, a post-test interface started and the participant answered the questions in the test. An interview was conducted after the tutorial in order to measure the participant satisfaction and their opinion about using LANA-II. During the tutorial and the interview, the student was observed by the researcher and the teacher to notice the participant's behaviors, facial expressions, and the interactions.

7.7.6 Case Study: Results and discussion

In this section, the data that was gathered from the interview (Table 7.39, Table 7.40) and the observations are described and classified in order to answer the question related to Hypothesis C (LANA-II is appropriate as a virtual tutor for Autistic students).

Question: Are the Autistic students satisfied with the usability of LANA-II?

- **Mariam:** The first participant interviewed in this study was, Mariam. She is eleven years old in Grade 5 and she has an early community clinical diagnosis of Asperger's, with no problem with language and intellectual impairments. Based on the teacher's interview, Mariam made an improvement in her behaviour during her study in the school from Grade 1 to 5. Her education level is very good especially in Maths and English. She is interested in using technology such as a computer, iPad, and mobile devices.

The study begun by detecting Mariam's learning style using the VAK questionnaire. The results from the questionnaire showed that her learning style was Visual. Mariam completed the pre-test and then interacted with the tutorial and

answered the questions. During the tutorial, videos/images appeared for each information in the tutorial. After she completed the tutorial, a post-test interface started and she answered the questions in the post-test. Mariam enjoyed using LANA-II as indicated in her interview answers. She said that she especially liked watching to the videos and the pictures during the tutorial. She was able to use LANA-II independently with ease and required no additional support once told how to use it. Her learning was improved by measuring learning gain and the relative learning gain between the pre-test and the post-test scores. Table 7.41 shows the pre-test and the post-test scores for all three participants. She was able to remember what she had learned from LANA-II, when she was asked in the interview. The rest of the interview questions were positive but when she was asking if she like learning from LANA-II more than learning from her teacher, she answered 'Maybe'. The teacher said that Mariam interacted positively with LANA-II and her interaction with LANA-II was better than in the classroom. According to the teacher's comments, Mariam does not like to listen to the lesson in the classroom for more than 15 minutes, she feels bored and she starts to do another task. Whereas, during the experiment, Mariam spent 25 minutes and she did not exhibit her typical behaviour associated with boredom and she completed the whole tutorial.

- **Fatimah:** The second participant interviewed in this study, Fatimah. She is eleven years old in Grade 5 and she had a late diagnosis of Asperger's, when she was nine years old. Based on the teacher's interview, Fatimah's education level is below the average because she does not like to study although she has an excellent memory. She has the ability to remember everything she have seen or read. Her learning style was detected using the questionnaire and it was found that her learning style was Visual. Fatimah completed the pre-test and then interacted with the tutorial and answered the questions. When she completed the tutorial, a post-test interface started and she answered the questions in the test. Fatimah said that using LANA-II was very enjoyable. She liked watching to the videos and the pictures during the tutorial. She was using LANA-II independently and easily with no additional support once she was told in how to use it. According to her scores in pre-test and post-test (Table 7.41), her learning gain the relative learning gain were improved the post-test score was greater than the pre-test score. She was able to remember what she have been learned from LANA-II, when she was asked in the interview.

When she was asking if she like learning from LANA-II more than learning from her teacher, she answered 'Both'. The rest of the questions were answered positively. The teacher said that Fatimah interacted with LANA-II better than in the classroom. Based on the teacher comments, Fatimah feels sleepy in the classroom and sometimes she sleeps during the lessons. On the other hand, during the experiment Fatimah spent 30 minutes, she did not feel sleepy and she completed the whole tutorial.

- **Dana:** The third participant interviewed in this study, Dana, she is ten years old in Grade 4 and she also had a late diagnosis of Asperger's, when she was nine years old. Based on the teacher's interview, Dana's education level is above the average. She is very good student in all subjects. Her learning style was detected using the questionnaire and it was found that her learning style was Kinaesthetic. Dana completed the pre-test and then interacted with the tutorial and answered the questions. When she completed the tutorial, a post-test interface started and she answered the questions in the test. Dana was very interested when using LANA-II CITS and during the tutorial she was saying comments with positive words such as 'nice', 'I like it', 'I know the answer'. She was very excited when she used the magnetism model and the electric circuit during the tutorial. She was using LANA-II CITS independently and easily with no additional support once she was told in how to use it. She made a good improvement on her learning gain and the relative learning gain based on her scores in pre-test and post-test (Table 7.41). The post-test score was greater than the pre-test score. When she was asked in the interview if she can remember the tutorial, she said that she remembered it and she recalled information about the tutorial. She said that she preferred to learn from LANA-II more than from her teacher. The rest of the questions were answered positively. The teacher said that Dana also interacted with LANA-II better than in the classroom because Dana likes learning from the computer more than learning from the teacher.

Table 7.41: The pre-test and post-test scores for the three participants

Student's Name	Pre-Test	Post-Test	Learning gain	Relative learning gain
Mariam	4	9	5	80%
Fatimah	5	10	5	100%
Dana	6	10	4	100%

7.8 Discussion

When looking at experiment 2 with the autistic students, the evaluation results taken from the log file reveal that the post-test scores are ranked higher, which means that the students performed well in the post-test. The P-values less than 0.05 ($p = .001$) (Section 7.5.3.1), indicates there was a statistically significant difference between the pre-test and post-test scores when adapting to the VAK learning style model. In addition, the learning gain between the Control Group (without adapting to VAK) and the Experimental Group (with adapting to VAK) were also statistically significant P-value less than 0.05 ($P = .000009$). In addition, the Experimental Group had a higher mean rank than the Control Group, meaning that the leaning gain for the Experimental Group was significantly improved compared to the learning gain of the Control Group (Section 7.5.3.1). The results from the questionnaire reveal the participant's perception of remembering what LANA-II has taught them between the Control Group and the Experimental Group. The majority of students in the Experimental Group reported remembering the tutorial 93.3% and 6.7% of the students had a neutral feeling; whereas students in the Control Group reported remembering the tutorial with a percentage of 53.3% and 46.7% stated they had a neutral feeling when rating this question (Section 7.5.3.2). The analysis of results showed that there was a statistically significant relationship between adapting or not adapting to the student's learning style and answering this question P- value less than .05 ($P = .013$), indicating that adapting to VAK learning style helped the students to remember the tutorial. Furthermore, the evaluation results taken from the questionnaire reveal the participant's perception of adapting their learning style in the tutorial. The majority of the students 93.3% were happy with adapting their learning style, whereas 6.7% of the students had a neutral

feeling when rating this question (Section 7.5.3.2). When looking at the results of the participant's preference of learning from LANA-II rather than learning from the teacher, the evaluation results reveal that the majority of students 93.3% in the Experimental Group preferred to learn from LANA-II more than learning from the teacher, and 6.7% of the students had a neutral feeling. The students in the Control Group were preferring to learn from LANA-II than learning from the teacher with percentage of 60%, and 40% stated they had a neutral feeling when rating this question. In addition, there was statistically significant relationship between adapting or not adapting to the learning style and answering this question P- value less than .05 ($P=.031$) (Section 7.5.3.2). It can be concluded from the above results and analysis that adapting LANA-II to autistic students using the VAK model succeeded. Therefore, the hypothesis **HA₁ (LANA-II with VAK model can be adapted to the Autistics learning style)** can be accepted.

When looking at the improvements made to the LANA-II algorithm, a morphologic pre-processing method was applied to address the common spelling variations and mistakes made by user because of the phonological similarity of some of Arabic alphabet characters (see chapter 6 section 6.2). Furthermore, the similarity algorithm (i.e. Cosine similarity) was redeveloped to allow the common variations of certain words to be recognised and responded to correctly (see chapter 6 section 6.3). These improvements have reduced the number of unrecognised utterances and incorrect responses which were evident in the results. The evaluation results taken from the log file revealed that unrecognised utterances were reduced to 5.4% in LANA-II (Chapter 7, Table 7.21) compared to 38% in LANA-I (Chapter 5, Table 5.19). The results of the test revealed a statistically significant difference between the numbers of unrecognised utterances, with LANA-II having a lower mean rank of unrecognised utterances. The results also showed that the incorrect responses were reduced to 3.6% (Table 7.21) in the LANA-II compared to 10.2% in LANA-I (Chapter 5, Table 5.19). There were statistically significant differences between the numbers of the incorrect response between LANA-I and LANA-II ($p = 0.001$). In terms of the correct responses, the results revealed that the correct responses using STS algorithm in LANA-II was significantly improved 96% compared to the correct responses using STS algorithm of LANA-I 89%. The findings of the questionnaire from the evaluation of LANA-I and LANA-II showed there was a significant difference in question 8 (the

naturalness) $p = 0.000163$. This revealed that the users perceived the LANA-II to be better in relation to conversation naturalness with 91.6% of the participants expressing that the conversation level of naturalness was good which is a major improvement from LANA where 45.8% who expressed the conversation level of naturalness was good. Hence based on these results there is enough evidence to suggest that **HB₁ (The enhancements made in LANA-II architecture improve the overall effectiveness of LANA-II engine)** can be accepted.

Additional findings of interest derived from the evaluation of LANA-II were highlighted through the comparison of the data gathered from the autistic students and neurotypical students. The results show that there was a statistically significant difference in the number of unrecognized utterance between the two participants ($p = .045$). The neurotypical participants mean values rank higher than the autistic participants, indicating that the number of unrecognised utterances from neurotypical participants was more than the autistic participants. In addition, the number of correct responses between the two participants was statistically significant ($p = .010$). The autistic participants had more correct responses than the neurotypical participants. Furthermore, the results indicated that there were statistically significant differences between the numbers of the incorrect response between the autistic participants and the neurotypical participants ($p = 0.009$), meaning the neurotypical participants had more incorrect responses than the autistic participants. Most prominent of the differences observed between the autistic students and the neurotypical students was in terms of the number of utterances that used the new STS algorithm, the results revealed that the number of utterances required the new STS algorithm in neurotypical participants 42.7% were more than autistic participants 31.12%. The comparison of the data from the two participant datasets illustrated that the utterances from the neurotypical participants contained significantly more instances where the utterance required the new STS algorithm. Most of the utterances from the autistic participants were matched the scripted patterns without need to apply the STS algorithm. A reason for this could be that the way that the autistic students speak and write. They tend to use a classical to modern Arabic language (formal language) on their speaking and writing, whereas the neurotypical students usually speak and write modern to colloquial Arabic language. (unformal language). The language used in the scripted patterns is modern Arabic language (formal language), therefore the autistic student's

utterances matched the scripted patterns, whereas the neurotypical students used some colloquial language in their writing, therefore their utterances didn't match the scripted patterns and required the string similarity algorithm.

The evaluation results taken from the case study of three autistic students showed an improvement on the autistic students learning gain and the relative learning gain when comparing their scores in the pre-test and the post-test (Table 7.41). Observation and interview questions, which were answered by both the student and the teacher, were used to collect the data. In terms of the student preference of learning from LANA-II more than learning from the teacher, the results show few differences in their answers. One student said that she liked the learning from LANA-II more than the learning from the teacher, whereas the two students were hesitated when they answered the question and one of them answered 'Maybe' and the other said 'Both'. A reason for this could be that the teacher presence during the interview. In terms of the teacher interview results, the teacher particularly liked the adapting to the student's learning style in the tutorial because this will meet the individual learning needs for the students with ASD. Furthermore, the teacher reported positive comments with regards the student's interaction with the tutorial using LANA-II comparing with the classroom. The teacher reported that one student feels sleepy and the other feels bored all the time in the classroom, whereas when they used LANA-II they were very interested and excited, because many students with ASD enjoy the use of technology. The teacher recommended that using LANA-II in the classroom would be beneficial because it will increase the students' independence from the teacher. For example, students with using LANA-II will not need to wait for a teacher to explain the lesson and then the teacher can focus in other aspects such as the student's behaviour, and social skills. Based on these results, it can be concluded that there is enough evidence to support **HC₁ (LANA CITS is appropriate as a virtual tutor for Autistic students).**

Inevitably, as with any research and evaluation effort the evaluation of LANA-II did highlight some areas that can be improve through further research. It is found that the knowledge base has some gaps in its domain and general knowledge which cause some unrecognised utterances. However, it is easy to address this weakness by adding the missing information (scripted patterns) into the knowledge base. It is expected to have this type of weakness because it is hard to cover all questions and general knowledge

that could be asked by the user. This is discussed in detail in the further work, chapter 8.

7.9 Chapter Summary

This chapter has detailed the experimental methodology that used to evaluate LANA-II through three experiments. The success of LANA-II was evaluated by measuring the success of three hypotheses:

1. Hypothesis A: LANA-II can be adapted to the autistic students learning style.
2. Hypothesis B: The enhancements made to LANA-II architecture improve the overall effectiveness of LANA-II engine.
3. Hypothesis C: LANA-II is appropriate as a virtual tutor for Autistic students.

The results from the experiments, using a combination of subjective and objective measures, show a significant evidence to support all the hypotheses. The implications of these findings and results on the research hypothesis are discussed in the following chapter.

The main novel contributions in this chapter are:

- Conducted an experiment to test the enhancement made in LANA-II engine by comparing the metrics from the (LANA-I) with the same metrics from LANAI.
- Evaluated LANA-II by conducting an experiment with Autistic children to measure the capabilities from different aspects such as tutoring success, adapting to VAK learning style model and user evaluation.
- Comparative descriptive analysis of data from autistic participants and neurotypical participants.
- A case study evaluation with three autistic students to test if LANA-II is appropriate as a virtual tutor for Autistic students.

Chapter 8 – Conclusion and Future Work

8.1 Overview

The research presented in this thesis aimed to answer two research questions:

1. Can an Arabic Conversational Intelligent Tutoring System (CITS) be adapted to the learning styles for autistic children?
2. Can an Arabic Conversational Intelligent Tutoring System (CITS) for children with autism enhance their learning?

To answer these questions an Arabic Conversational Intelligent Tutoring System for autistic children (LANA) was researched, designed and developed. The research involved investigation in to several key areas such as, Conversational Agents (CAs), Intelligent tutoring systems (ITS), Arabic Language Processing techniques (i.e. natural language processing, sentence similarities measures and pattern matching), the VAK learning style, and the TEACCH method for children with autism spectrum disorder. Two prototypes were developed in this research. The first prototype was (LANA-I) that contains two main components: the first component of LANA-I is the CA engine. Due to grammatical features and the morphological nature of the Arabic language, the research into CA development techniques revealed that a hybrid approach was the most appropriate approach to develop an Arabic CA. The LANA-I CA engine is based on the two main CA development strategies, pattern matching (PM) and short text similarity (STS) that calculates the matching strength of a pattern to the user utterance. The two parts of the engine work together in order to overcome some of the unique language challenges of the Arabic language. PM is a popular method for developing CAs, however CAs based on the PM principle have disadvantages such as the large number of scripted patterns that have to be covered in order to create a robust knowledge base specially with Arabic language due to the free word order in the Arabic language. The second component of LANA-I is the intelligent tutoring system which adapts the tutorial to the autistic children learning style. LANA-I personalises the tutorial according to the autistic children learning style: Visual, Auditory, Kinaesthetic (VAK). The results of the log file and end user evaluation of LANA-I revealed some weaknesses and some negative perceptions from the participants. The issues that were highlighted in LANA-I evaluation were mainly due to its grammatical features and the morphological nature. Other issues that were identified were the need

to further expand the knowledge base in order to increase the naturalness of the conversation. The participants expressed that they perceived the naturalness of their conversation with LANA-I to be low.

The second prototype LANA-II aimed to overcome these weaknesses. LANA-II CA algorithm was developed in order to reduce the effort required in scripting the knowledge base/domain by using pre-processing feature and new version of Cosine similarity approach. In addition, the tutorial interface is mapped to the TEACCH method, which is based on Structured Teaching technique for autistic children. From the evaluation of LANA-II, statistically significant improvements were observed in terms of the objective and subjective metrics measured in LANA-II CA engine. Furthermore, LANA-II using VAK learning style succeeded in adapting to the autistic students learning style and enhance their learning based on the objectives and subjective metrics from the experiment and the results from the case study. Additional findings of interest derived from the evaluation of LANA-II were highlighted through the comparison of the data gathered from the autistic students and neurotypical students. The results show differences observed between the autistic students and the neurotypical students in terms of the number of utterances using the STS algorithm. The results revealed that the number of utterances that used the STS algorithm in neurotypical participants was higher than autistic participants. Most utterances from the autistic participants matched the scripted patterns without need to apply the STS algorithm. A reason for this could be that the way that the autistic students speak and write; they tend to use a classical to modern Arabic language (formal language) on their speaking and writing, whereas neurotypical students usually speak and write modern to colloquial Arabic language (informal language). The language used in the scripted patterns is modern Arabic language (formal language), therefore the autistic student's utterances matched the scripted patterns, whereas the neurotypical students used some colloquial language in their writing, therefore their utterances did not match the scripted patterns and required the string similarity algorithm.

Based on the results observed from the evaluations, the evidence supports the conclusion that LANA-II can be adapted to autistic students learning style and enhance their learning. In addition, the LANA-II CA engine is effective as a CA in Arabic language. Further work in the field of CITS development can build a top of this work which is discussed in section 8.3.

8.2 Research Contribution

This research has produced some significant contributions in the field of conversational intelligent tutoring system. The primary aim of this research was to design and implement an Arabic CITS for high functioning autistic students called LANA using the VAK model to adapt to autistic learning style and enhance their learning in mainstream schools. LANA architecture uses the pattern matching (PM) and an Arabic short text similarity (STS) measure to extract facts from user's responses to match rules in scripted conversation in a particular domain (Science). The outcome of the project is the development of an Arabic CITS for autistic students (10 to 12 years old) who have previously received a community clinical diagnosis of high functioning autism spectrum disorder or Asperger's, who they have no language problem and intellectual impairment. Also, they are studying in Arabic mainstream schools.

The prominent contributions derived from this research are as follows:

8.2.1 Method for developing LANA CA engine

The research has led to the development of a novel CA engine based on hybrid similarity method, which contains pattern matching principles and lexical string similarity in order to converse with the user to reach the goal of the conversation. The LANA CA engine handles the language unique features such as the grammatical features and the morphological nature of the Arabic language. Thus, this research makes a contribution in terms of a framework for developing CA engines in Arabic language. Pre-processing methods were applied on the utterance before applying the similarity algorithm in order to overcome the spelling mistake issues due to the phonological similarity of some of its alphabet characters. In addition, the cosine similarity was adjusted to use the character level approach by comparing each word with every other word character by character after applying the Bi-gram method. This algorithm solves the complex word order issue that comes with the Arabic language. It also significantly reduces the number of scripts that have to be written to deal with the issue of word order or word structure, such as adding the prefix or suffix.

8.2.2 Method for developing an Arabic scripting language for LANA CITS

A new Arabic scripting language has been developed in order to script the domain. The domain used to script the CA was from the Science curriculum for the age 10-12 years old. The material was provided by the Ministry of Education in Saudi Arabia. Scripts consisting of a context structured according to the topics in the science book such as the Earth, Moon, Solar System, eclipse, etc.). LANA CITS is designed to deliver the Science topic to Autistic learners in the Arabic language based on the science book that is taught in schools in Saudi Arabia. The scripting language proposed in this research contains new features that deal with the unique features of the language such as allowing word order variations of certain patterns. In addition, the scripting language includes supporting material to the user depending on the user's learning style (Visual: images and videos – Auditory: Sounds – Kinaesthetic: Instructions and objects).

8.2.3 Method for adapting the tutorial to the VAK learning style

A proof of concept adaptation the tutorial to the autistic VAK learning style has been researched, developed and implemented into LANA CITS. LANA CITS has the ability to personalise the lesson to the user's learning style (VAK) and provide suitable material to the user according to the user's learning style (images and videos for Visual learner, Sound for Auditory learner, physical models for Kinaesthetic learner). Adapting the tutorial to the autistics' VAK learning style was evaluated and it significantly improve the students' learning gain and enhance their learning.

8.2.4 Method for designing GUI using TEACCH method

A new GUI was designed, implemented and evaluated that adopted the TEACCH method, which is an educational approach for people with ASD. In this research the TEACCH method (G. Mesibov, 2018) WAS adopted in GUI by applying the four main TEACCH components: physical structure, Daily schedules, Work system, and Visual cues and instructions (Chapter 6, Section 6.5). LANA is the first CITS that mapped TEACCH method to the user interface.

8.2.5 Framework for CITS evaluation

A new evaluation framework has been researched and tested in order to give an overall performance related CITS evaluation. The proposed methodology focuses on evaluating metrics related to the CITS ability to achieve the goal of its development by employing software evaluation methodologies such as the Goal Question Metric (GQM). This mechanism was based on supposition that goals must be identified for a system then number of questions was defined for each goal, and a set of metrics associated with each question (Fenton and Pfleeger, 1998). In addition, a qualitative evaluation was used by conducting three case studies in order to test the system in depth from the autistic children point of view and to notice their interactions and perceptions during the tutorial.

8.3 Future Work

The research presented in this thesis has outlined a novel approach to Arabic conversation intelligent tutoring system that adapts to autistic children learning style and enhance their learning. The conversation engine was designed in an Arabic language which is resource poor and completely different to English in its grammatical and morphological structure. However, there are areas which can be improved through further research and development such as the suggestions below:

8.3.1 Semantic similarity

LANA CA similarity algorithm can be strengthened with the addition of semantic rather than lexical similarity when a suitable Arabic WordNet is available. The paraphrased version of a scripted pattern will be recognised by using the semantic similarity. This will further reduce the scripting patterns and making the task of scripting an Arabic CA even less exhausting. Future research could be made to the Arabic WordNet in order to extend its coverage and turn it into a more useful knowledge base.

8.3.2 Knowledge base expansion

As the knowledge base of LANA is based on the science subjects, specifically two lessons from the science book of grade 4, future work can entail the expansion of the knowledge base to cover all lessons of the science topics for grade 4. Moreover, the

knowledge base could be expanded to cover other lessons for other grades. Furthermore, the general topics of the knowledge base could also be expanded to include more topics and patterns related to general topics.

8.3.3 Automatically predict autistic children learning styles

As detecting the learning style in LANA is based on a questionnaire. However, if the questionnaire is not completed accurately, this leading to incorrect assessment of learning styles. Future work can be made to detect the autistic children learning style by analysing learner behaviour throughout tutoring.

8.3.4 Voice Recognition

Adding voice recognition feature will make LANA more fixable for autistic students as some individuals may prefer speaking to typing, whereas others may have low auditory tolerance and prefer typing.

References

- Abdel Ghafour, H. H., El-Bastawissy, A. and Heggazy, A. F. A. (2011) *AEDA: Arabic edit distance algorithm Towards a new approach for Arabic name matching*. IEEE.
- Abderrahim, M. A., Abderrahim, M. E. A. and Chikh, M. A. (2013) 'Using Arabic wordnet for semantic indexation in information retrieval system.' *arXiv preprint arXiv:1306.2499*,
- Abouenour, L., Bouzoubaa, K. and Rosso, P. (2010) *Using the Yago ontology as a resource for the enrichment of Named Entities in Arabic WordNet*.
- Abuleil, S., Alsamara, K. and Evens, M. (2002) *Acquisition system for Arabic noun morphology*. Association for Computational Linguistics.
- Ahuja, N. J. and Sille, R. (2013) 'A critical review of development of intelligent tutoring systems: Retrospect, present and prospect.' *International Journal of Computer Science Issues*, 10(4) pp. 39-48.
- Akplotsyi, R. and Mahdjoubi, L. (2011) 'Effects of learning styles on engaging children in school projects.'
- Al-marsoomi, F. (2015) *The Development of a Framework for Semantic Similarity Measures for the Arabic Language*. PhD. Manchester Metropolitan University.
- Al-Ramahi, M. A. and Mustafa, S. H. (2012) 'N-Gram-Based Techniques for Arabic Text Document Matching; Case Study: Courses Accreditation.'
- Al-Saidat, E. and Al-Momani, I. (2010) 'Future markers in Modern Standard Arabic and Jordanian Arabic: a contrastive study.' *European Journal of Social Sciences*, 12(3) pp. 397-408.
- Al-Thubaity, A. O. (2015) 'A 700M+ Arabic corpus: KACST Arabic corpus design and construction.' *Language Resources and Evaluation*, 49(3) pp. 721-751.
- Al-Zoghby, A. M., Ahmed, A. S. E. and Hamza, T. T. (2013) 'Arabic Semantic Web Applications—A Survey.' *Journal of Emerging Technologies in Web Intelligence*, 5(1) pp. 52-69.
- AlAgha, I. and Abu-Taha, A. (2015) 'AR2SPARQL: An Arabic Natural Language Interface for the Semantic Web.' *International Journal of Computer Applications*, 125(6)
- Alanazi, S., Sharp, B. and Stanier, C. (2015) 'A Named Entity Recognition System Applied to Arabic Text in the Medical Domain.' *International Journal of Computer Science Issues (IJCSI)*, 12(3) p. 109.
- Aljameel, S. S., O'Shea, J. D., Crockett, K. A. and Latham, A. (2016) *A Review of Current Technology-Based Intervention for School Aged Children with Autism Spectrum Disorder*. Springer.
- Alkhalifa, M. and Rodríguez, H. (2010) 'Automatically extending named entities coverage of Arabic WordNet using Wikipedia.' *International Journal on Information and Communication Technologies*, 3(3)

- Allen, J. F., Byron, D. K., Dzikovska, M., Ferguson, G., Galescu, L. and Stent, A. (2001) 'Toward conversational human-computer interaction.' *AI magazine*, 22(4) p. 27.
- Alobaidi, O. G., Crockett, K. A., O'Shea, J. D. and Jarad, T. M. (2013) *Abdullah: An Intelligent Arabic Conversational Tutoring System for Modern Islamic Education*. Vol. 2:
- Alqahtani, M. M. A. and Atwell, E. (2015) 'A Review of Semantic Search Methods to Retrieve Information from the Qur'an Corpus.'
- AmericanPsychiatricAssociation. (2013) 'Diagnostic and statistical manual of mental disorders.' *Arlington: American Psychiatric Publishing*,
- Banire, B., Jomhari, N. and Ahmad, R. (2015) 'Visual Hybrid Development Learning System (VHDLS) Framework for Children with Autism.' *Journal of Autism and Developmental Disorders*, 45(10) pp. 3069-3084.
- Barnard, J., Prior, A. and Potter, D. (2000) 'Autism and inclusion: Is it working.' *London: National Autistic Society*,
- Barnhart, R. (2017) *A Phenomenological Study of Elementary Teachers of Autism Spectrum Disorder Students: Common Lived Experiences*. University of Louisiana at Lafayette.
- Batet, M., Sánchez, D. and Valls, A. (2011) 'An ontology-based measure to compute semantic similarity in biomedicine.' *Journal of biomedical informatics*, 44(1) pp. 118-125.
- Bauminger-Zviely, N., Eden, S., Zancanaro, M., Weiss, P. L. and Gal, E. (2013) 'Increasing social engagement in children with high-functioning autism spectrum disorder using collaborative technologies in the school environment.' *Autism*, 17(3) pp. 317-339.
- Belinkov, Y., Habash, N., Kilgarriff, A., Ordan, N., Roth, R. and Suchomel, V. (2013) 'arTen-Ten: a new, vast corpus for Arabic.' *Proceedings of WACL*,
- Bellegarda, J. R. (2014) 'Spoken language understanding for natural interaction: The siri experience.' *In Natural Interaction with Robots, Knowbots and Smartphones*. Springer, pp. 3-14.
- Berg, B. L. (2004) *Methods for the social sciences*. Pearson Education Inc, United States of America.
- Bernard-Opitz, V., Sriram, N. and Nakhoda-Sapuan, S. (2001) 'Enhancing Social Problem Solving in Children with Autism and Normal Children Through Computer-Assisted Instruction.' *Journal of Autism and Developmental Disorders*, 31(4) p. 377.
- Bickmore, T. and Giorgino, T. (2006) 'Health dialog systems for patients and consumers.' *Journal of biomedical informatics*, 39(5) pp. 556-571.
- BNC-Consortium. (2003) *British national corpus*.: Oxford-University. [Online] [Accessed <http://www.natcorp.ox.ac.uk/>].

- Boucenna, S., Narzisi, A., Tilmont, E., Muratori, F., Pioggia, G., Cohen, D. and Chetouani, M. (2014) 'Interactive technologies for autistic children: A review.' *Cognitive Computation*, 6(4) pp. 722-740.
- Boudabous, M. M., Kammoun, N. C., Khedher, N., Belguith, L. H. and Sadat, F. (2013) *Arabic WordNet semantic relations enrichment through morpho-lexical patterns*. IEEE.
- Bozgeyikli, L., Raij, A., Katkoori, S. and Alqasemi, R. (2018) 'A Survey on Virtual Reality for Individuals with Autism Spectrum Disorder: Design Considerations.' *IEEE Transactions on Learning Technologies*, 11(2) pp. 133-151.
- Budanitsky, A. and Hirst, G. (2001) *Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures*. Vol. 2:
- Burkard, R. E. and Cela, E. (1999) 'Linear assignment problems and extensions.' *In Handbook of combinatorial optimization*. Springer, pp. 75-149.
- Butler, K. A. (1987) *Learning and teaching style: In theory and practice*. Learners Dimension.
- Byrne, B. (2004) 'Qualitative interviewing.' *Researching society and culture*, 2 pp. 179-192.
- Cagliani, R. R., Ayres, K. M., Whiteside, E. and Ringdahl, J. E. (2017) 'Picture exchange communication system and delay to reinforcement.' *Journal of Developmental and Physical Disabilities*, 29(6) pp. 925-939.
- Campbell, A. (2016) *TECHNOLOGY ASSISTED SEXUAL HEALTH AGENT (TASHA): A CONVERSATIONAL AGENT FOR PATIENT DECISION MAKING ON THE MANAGEMENT OF CERVICAL DYSPLASIA/CANCER*. Smdm.
- Cassell, J. (2000) *Embodied conversational agents*. MIT press.
- Coffield, F., Moseley, D., Hall, E. and Ecclestone, K. (2004) 'Learning styles and pedagogy in post-16 learning: A systematic and critical review (application/pdf Object).'
- Cohen, J. (1988) *Statistical power analysis for the behavioral sciences* 2nd edn. Erlbaum Associates, Hillsdale.
- Corbetta, P. (2003) *Social research: Theory, methods and techniques*. Sage.
- Creswell, J. W., Hanson, W. E., Clark Plano, V. L. and Morales, A. (2007) 'Qualitative research designs: Selection and implementation.' *The counseling psychologist*, 35(2) pp. 236-264.
- Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research*. Sage publications.
- Crockett, K., James, O. S. and Bandar, Z. (2011a) *Goal orientated conversational agents: applications to benefit society*. Springer.
- Crockett, K., James, O. S. and Bandar, Z. (2011b) 'Goal orientated conversational agents: applications to benefit society.' *In Agent and Multi-Agent Systems: Technologies and Applications*. Springer, pp. 16-25.

- Crockett, K., Latham, A. and Whitton, N. (2017) 'On predicting learning styles in conversational intelligent tutoring systems using fuzzy decision trees.' *International Journal of Human-Computer Studies*, 97 pp. 98-115.
- Crockett, K., Bandar, Z., O'Shea, J. and McLean, D. (2010) *Goal Orientated Conversational Agents—The rocky road to commercialization*. IEEE.
- Crow, T., Luxton-Reilly, A. and Wuensche, B. (2018) *Intelligent tutoring systems for programming education: a systematic review*. ACM.
- Crystal, D. (2008) 'Front Matter.' *A Dictionary of Linguistics and Phonetics, Sixth Edition*, pp. i-xxv.
- Dasgupta, D., Hernandez, G., Garrett, D., Vejandla, P. K., Kaushal, A., Yerneni, R. and Simien, J. (2008) *A comparison of multiobjective evolutionary algorithms with informed initialization and kuhn-munkres algorithm for the sailor assignment problem*. ACM.
- Davis, P., Florian, L. and Ainscow, M. (2004) *Teaching strategies and approaches for pupils with special educational needs: A scoping study*. DfES Publications.
- Demir, Ş. (2003) *Improved treatment of word meaning in a Turkish conversational agent*. Citeseer.
- Dermeval, D., Paiva, R., Bittencourt, I. I., Vassileva, J. and Borges, D. (2017) 'Authoring Tools for Designing Intelligent Tutoring Systems: a Systematic Review of the Literature.' *International Journal of Artificial Intelligence in Education*, pp. 1-49.
- Dunn, R. and Dunn, K. (1974) 'Learning style as a criterion for placement in alternative programs.' *The Phi Delta Kappan*, 56(4) pp. 275-278.
- Durlak, J. A. (2009) 'How to select, calculate, and interpret effect sizes.' *Journal of pediatric psychology*, 34(9) pp. 917-928.
- D'Elia, L., Valeri, G., Sonnino, F., Fontana, I., Mammone, A. and Vicari, S. (2014) 'A longitudinal study of the TEACCH program in different settings: The potential benefits of low intensity intervention in preschool children with autism spectrum disorder.' *Journal of autism and developmental disorders*, 44(3) pp. 615-626.
- Edelson, S. M. (2000) 'Learning styles and autism.' *Center for the Study of Autism, Salem, Oregon*. En <http://www.autism.org/styles.html>,
- Education, M. o. (2017) *Ministry of Education*. [Online] [Accessed <https://www.moe.gov.sa/en/Pages/default.aspx>]
- Elmagarmid, A. K., Ipeirotis, P. G. and Verykios, V. S. (2007) 'Duplicate record detection: A survey.' *IEEE Transactions on knowledge and data engineering*, 19(1) pp. 1-16.
- Felder, R. M. and Silverman, L. K. (1988) 'Learning and teaching styles in engineering education.' *Engineering education*, 78(7) pp. 674-681.
- Fenton, N. E. and Pfleeger, S. L. (1998) *Software Metrics: A Rigorous and Practical Approach*. Brooks. Cole.

- Fernezeyi, M., Maszák, Z. S. and Langh, R. (2006) *Smalltalk: interactive installation*. ACM.
- Fleming, N. D. (2001) *Teaching and learning styles: VARK strategies*. IGI Global.
- Fletcher-Watson, S. (2014) 'A targeted review of computer-assisted learning for people with autism spectrum disorder: Towards a consistent methodology.' *Review Journal of Autism and Developmental Disorders*, 1(2) pp. 87-100.
- GlobalWordNet. (2018) <http://globalwordnet.org/wordnets-in-the-world>. [Online] [Accessed <http://globalwordnet.org/wordnets-in-the-world>]
- Gomaa, W. H. and Fahmy, A. A. (2012) 'Short answer grading using string similarity and corpus-based similarity.' *International Journal of Advanced Computer Science and Applications (IJACSA)*, 3(11)
- Gomaa, W. H. and Fahmy, A. A. (2013) 'A survey of text similarity approaches.' *International Journal of Computer Applications*, 68(13)
- Graesser, A. C., Chipman, P., Haynes, B. C. and Olney, A. (2005) 'AutoTutor: An intelligent tutoring system with mixed-initiative dialogue.' *IEEE Transactions on Education*, 48(4) pp. 612-618.
- Gregor, E. M. C. and Campbell, E. (2001) 'The attitudes of teachers in Scotland to the integration of children with autism into mainstream schools.' *Autism*, 5(2) pp. 189-207.
- Grynszpan, O., Martin, J.-C. and Nadel, J. (2008) 'Multimedia interfaces for users with high functioning autism: An empirical investigation.' *International Journal of Human - Computer Studies*, 66(8) pp. 628-639.
- Grynszpan, O., Weiss, P. L., Perez-Diaz, F. and Gal, E. (2014) 'Innovative technology-based interventions for autism spectrum disorders: a meta-analysis.' *Autism*, 18(4) pp. 346-361.
- Gutiérrez, Y., Fernández, A., Montoyo, A. and Vázquez, S. (2010) *UMCC-DLSI: Integrative resource for disambiguation task*. Association for Computational Linguistics.
- Habash, N. Y. (2010) 'Introduction to Arabic natural language processing.' *Synthesis Lectures on Human Language Technologies*, 3(1) pp. 1-187.
- Hajeer, I. (2012) 'Comparison on the Effectiveness of Different Statistical Similarity Measures.' *International Journal of Computer Applications*, 53(8)
- Hatzivassiloglou, V., Klavans, J. L. and Eskin, E. (1999) *Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning*. Citeseer.
- Herod, L. (2004) 'Learning styles and strategies.' *Adult Learning & Literacy*, pp. 310-800.
- Hijjawi, M., Bandar, Z. and Crockett, K. (2016) 'The Enhanced Arabchat: An Arabic Conversational Agent.' *International Journal of Advanced Computer Science and Applications*,

- Hijjawi, M., Bandar, Z., Crockett, K. and McLean, D. (2014) *ArabChat: an Arabic Conversational Agent*. IEEE.
- Holloway, I. and Galvin, K. (2016) *Qualitative research in nursing and healthcare*. John Wiley & Sons.
- Hone, K. S. and Graham, R. (2000) 'Towards a tool for the subjective assessment of speech system interfaces (SASSI).' *Natural Language Engineering*, 6(3&4) pp. 287-303.
- Honey, P. and Mumford, A. (2006) 'The Learning Styles Questionnaire, 80-item version (Revised edition, July 2006).' *Peter Honey Publication*. [Online] last accessed December, 17 p. 2012.
- Howard-Jones, P. A. (2014) 'Neuroscience and education: myths and messages.' *Nature Reviews Neuroscience*, 15(12) p. 817.
- Islam, A. and Inkpen, D. (2008) 'Semantic text similarity using corpus-based word similarity and string similarity.' *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2(2) p. 10.
- Jee, B. D., Gentner, D., Uttal, D. H., Sageman, B., Forbus, K., Manduca, C. A., Ormand, C. J., Shipley, T. F. and Tikoff, B. (2014) 'Drawing on experience: how domain knowledge is reflected in sketches of scientific structures and processes.' *Research in Science Education*, 44(6) pp. 859-883.
- Kadri, Y. and Benyamina, A. (1992) *A syntax semantic analyzer for Arabic language*. University of Oran.
- Kaleem, M., O'Shea, J. and Crockett, K. (2014) 'Development of UMAIR the Urdu Conversational Agent for Customer Service.'
- Kaleem, M., Alobadi, O., O'Shea, J. and Crockett, K. (2016) *Framework for the Formulation of Metrics for Conversational Agent Evaluation*.
- Kara, S., Alan, Ö., Sabuncu, O., Akpınar, S., Cicekli, N. K. and Alpaslan, F. N. (2012) 'An ontology-based retrieval system using semantic indexing.' *Information Systems*, 37(4) pp. 294-305.
- Kelly, D. and Tangney, B. (2006) 'Adapting to intelligence profile in an adaptive educational system.' *Interacting with computers*, 18(3) pp. 385-409.
- Kolb, D. A. (1976) *Learning style inventory technical manual*. McBer Boston, MA.
- Kolb, P. (2008) 'Disco: A multilingual database of distributionally similar words.' *Proceedings of KONVENS-2008, Berlin*,
- Kondrak, G. (2005) *N-gram similarity and distance*. Springer.
- Kujala, T., Lepistö, T. and Näätänen, R. (2013) 'The neural basis of aberrant speech and audition in autism spectrum disorders.' *Neuroscience & Biobehavioral Reviews*, 37(4) pp. 697-704.
- Kulik, J. A. and Fletcher, J. D. (2016) 'Effectiveness of intelligent tutoring systems: a meta-analytic review.' *Review of Educational Research*, 86(1) pp. 42-78.

- Kvale, S. and Brinkmann, S. (2009) 'Learning the craft of qualitative research interviewing.' *Thousands Oaks: Sage Publications*,
- Landauer, T. K., Foltz, P. W. and Laham, D. (1998) 'An introduction to latent semantic analysis.' *Discourse processes*, 25(2-3) pp. 259-284.
- Latham, A., Crockett, K. A. and Bandar, Z. (2010) *A Conversational Expert System Supporting Bullying and Harassment Policies*. Citeseer.
- Latham, A., Crockett, K. and McLean, D. (2014) 'An adaptation algorithm for an intelligent natural language tutoring system.' *Computers & Education*, 71 pp. 97-110.
- Latham, A., Crockett, K., McLean, D. and Edmonds, B. (2012) 'A conversational intelligent tutoring system to automatically predict learning styles.' *Computers & Education*, 59(1) pp. 95-109.
- Latham, A. M. (2011) *Personalising learning with dynamic prediction and adaptation to learning styles in a conversational intelligent tutoring system*. Manchester Metropolitan University.
- Lee, J.-L., Orwig, G., Gunter, G. and Witta, L. (2006) *The effect of cognitive styles on a visually-oriented task in online learning environments*. Association for the Advancement of Computing in Education (AACE).
- Lester, J., Branting, K. and Mott, B. (2004) 'Conversational agents.' *The Practical Handbook of Internet Computing*, pp. 220-240.
- Li, Y., Bandar, Z., McLean, D. and O'Shea, J. (2004) *A Method for Measuring Sentence Similarity and its Application to Conversational Agents*.
- Li, Y., McLean, D., Bandar, Z. A., O'Shea, J. D. and Crockett, K. (2006) 'Sentence similarity based on semantic nets and corpus statistics.' *IEEE transactions on knowledge and data engineering*, 18(8) pp. 1138-1150.
- Limoges, E. I., Mottron, L. and Bolduc, C. B., Claude Godbout, Roger. (2005) 'Atypical sleep architecture and the autism phenotype.' *Brain*, 128 1061,
- Lindsay, S., Proulx, M., Scott, H. and Thomson, N. (2014) 'Exploring teachers' strategies for including children with autism spectrum disorder in mainstream classrooms.' *International Journal of Inclusive Education*, 18(2) pp. 101-122.
- Lisle, A. M. (2007) 'Assessing learning styles of adults with intellectual difficulties.' *Journal of Intellectual Disabilities*, 11(1) pp. 23-45.
- Loper, E. and Bird, S. (2002) *NLTK: The natural language toolkit*. Association for Computational Linguistics.
- Lord, C., Cook, E. H., Leventhal, B. L. and Amaral, D. G. (2013) 'Autism spectrum disorders.' *Autism: The Science of Mental Health*, 28 p. 217.
- Lutfi, S. L., Fernández-Martínez, F., Lorenzo-Trueba, J., Barra-Chicote, R. and Montero, J. M. (2013) 'I feel you: the design and evaluation of a domotic affect-sensitive spoken conversational agent.' *Sensors*, 13(8) pp. 10519-10538.

- Ma, W. (2017) *Intelligent Tutoring Systems and Learning Outcomes: Two Systematic Reviews*. Education: Faculty of Education.
- Mackay, A. (2010) *Motivation, ability and confidence building in people*. Routledge.
- Makrygianni, M. K., Gena, A., Katoudi, S. and Galanis, P. (2018) 'The effectiveness of applied behavior analytic interventions for children with Autism Spectrum Disorder: A meta-analytic study.' *Research in Autism Spectrum Disorders*, 51 pp. 18-31.
- Mandell, D., Novak, M. and Zubritsky, C. (2005) 'Factors Associated With Age of Diagnosis Among Children With Autism Spectrum Disorders.' *Pediatrics*, 116 1486,
- Maragoudakis, M., Kladis, B., Tsopanoglou, A., Sgarbas, K. and Fakotakis, N. (2001) *Natural language in dialogue systems. A case study on a medical application*.
- Marcus, A. and Wilder, D. (2009) 'A comparison of peer video modelling and self video modelling to teach textual responses in children with autism.' *Journal of applied behaviour analysis*, 42 341,
- Marcus, M. P., Marcinkiewicz, M. A. and Santorini, B. (1993) 'Building a large annotated corpus of English: The Penn Treebank.' *Computational linguistics*, 19(2) pp. 313-330.
- Marshall, C. and Rossman, G. (2006) 'The how of the study: Building the research design.' *Designing qualitative research*, pp. 55-101.
- Massaro, D. (2004) 'Symbiotic Value of an Embodied Agent in Language Learning.' *In Proceedings of the 37th Hawaii International Conference on System Sciences*: pp. 1-10.
- Massaro, D. W. and Bosseler, A. (2006) 'Read my lips: The importance of the face in a computer-animated tutor for vocabulary learning by children with autism.' *Autism*, 10(5) pp. 495-510.
- Meng, Q. (2017) Interventions promoting academic engagement for children with ASD in general education classrooms: A systematic literature review.
- Mesibov, G. (2018) *Accessing the curriculum for pupils with autistic spectrum disorders: Using the TEACCH programme to help inclusion*. Routledge.
- Mesibov, G. B. and Shea, V. (2010) 'The TEACCH program in the era of evidence-based practice.' *Journal of autism and developmental disorders*, 40(5) pp. 570-579.
- Mesibov, G. B., Howley, M. and NetLibrary, I. (2003) *Accessing the curriculum for pupils with autistic spectrum disorders: using the TEACCH programme to help inclusion*. London: David Fulton Publishers.
- Michie, D. and Sammut, C. (2001) 'Infochat scripter's manual.' *Convagent Ltd., Manchester*,

- Mihalcea, R., Corley, C. and Strapparava, C. (2006) *Corpus-based and knowledge-based measures of text semantic similarity*. Vol. 6:
- Miller, G. A. (1995) 'WordNet: a lexical database for English.' *Communications of the ACM*, 38(11) pp. 39-41.
- Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. J. (1990) 'Introduction to wordnet: An on-line lexical database*.' *International journal of lexicography*, 3(4) pp. 235-244.
- Miller, P. (2001) 'Learning Styles: The Multimedia of the Mind. Research Report.'
- Moore, D. S., McCabe, G. P. and Craig, B. A. (2012) 'Introduction to the Practice of Statistics.'
- Morik, K. (1991) 'Underlying assumptions of knowledge acquisition and machine learning.' *Knowledge Acquisition*, 3(2) pp. 137-156.
- Murray, D. (1997) *Autism and information technology: Therapy with computers*. In *Autism and Learning: A Guide to Good Practice*. Brookes.
- Musen, M. A. (1993) 'An overview of knowledge acquisition.' In *Second generation expert systems*. Springer, pp. 405-427.
- Noori, Z., Bandarl, Z. and Crockett, K. (2014) 'Arabic goal-oriented conversational agent based on pattern matching and knowledge trees.'
- Noori, Z., Crockett, K., Bandar, Z., & Al-Mousa, M. (2018, March). An Arabic Word Similarity Measure for Semantic Conversational Agents. In *2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)* (pp. 119-123). IEEE.
- O'Connell, C., Dance, R., Delasaux, V., Morling, E., Miller, S., Ranby, K., Russell, P., Stitt, C. and Stockman, K. (2013) *Supporting Children with Autistic Spectrum Disorders*. Routledge.
- O'Shea, J., Bandar, Z. and Crockett, K. (2011) *Systems Engineering and Conversational Agents*. Vol. 10 2011.
- O'Shea, K., Crockett, K. and Bandar, Z. (2010) *Application of a semantic-based conversational agent to student debt management*. IEEE.
- Organization, W. H. (2004) *International statistical classification of diseases and related health problems*. Vol. 1. World Health Organization.
- Owda, M., Bandar, Z. and Crockett, K. (2011) 'Information extraction for SQL query generation in the conversation-based interfaces to relational databases (C-BIRD).' In *Agent and Multi-Agent Systems: Technologies and Applications*. Springer, pp. 44-53.
- O'Roak, B. and State, M. (2008) 'Autism Genetics: Strategies, Challenges, and Opportunities.' *Autism Research*, 1 17,
- O'Shea, J., Bandar, Z., Crockett, K. and McLean, D. (2008) *A comparative study of two short text semantic similarity measures*. Springer.
- Padayachee, I. (2002) *Intelligent tutoring systems: Architecture and characteristics*.

- Panerai, S., Ferrante, L. and Zingale, M. (2002) 'Benefits of the Treatment and Education of Autistic and Communication Handicapped Children (TEACCH) programme as compared with a non-specific approach.' *Journal of intellectual disability research*, 46(4) pp. 318-327.
- Panerai, S., Zingale, M., Trubia, G., Finocchiaro, M., Zuccarello, R., Ferri, R. and Elia, M. (2009) 'Special education versus inclusive education: the role of the TEACCH program.' *Journal of autism and developmental disorders*, 39(6) pp. 874-882.
- Pannu, A. (2015) 'Artificial Intelligence and its Application in Different Areas.' *International Journal of Engineering and Innovative Technology*,
- Parkinson, D. B. (2012) ArabiCorpus.
- Parsons, S. and Cobb, S. (2011) 'State-of-the-art of virtual reality technologies for children on the autism spectrum.' *European Journal of Special Needs Education*, 26(3) pp. 355-366.
- Patton, M. Q. (2005) *Qualitative research*. Wiley Online Library.
- Pavlov, N. (2014) 'User interface for people with autism spectrum disorders.' *Journal of Software Engineering and Applications*, 7(02) p. 128.
- Perikos, I., Grivokostopoulou, F. and Hatzilygeroudis, I. (2017) 'Assistance and feedback mechanism in an intelligent tutoring system for teaching conversion of natural language into logic.' *International Journal of Artificial Intelligence in Education*, 27(3) pp. 475-514.
- Pietquin, O. and Hastie, H. (2013) 'A survey on metrics for the evaluation of user simulations.' *The knowledge engineering review*, 28(01) pp. 59-73.
- Power, Z., Campbell, M., Kilcoyne, P., Kitchener, H. and Waterman, H. (2010) 'The Hyperemesis Impact of Symptoms Questionnaire: development and validation of a clinical tool.' *International journal of nursing studies*, 47(1) pp. 67-77.
- Putnam, C. and Chong, L. (2008) 'Software and technologies designed for people with autism: what do users want?' In *10th International ACM SIGACCESS conference on computers and accessibility*: pp. 3-8.
- Qian, G., Sural, S., Gu, Y. and Pramanik, S. (2004) *Similarity between Euclidean and cosine angle distance for nearest neighbor queries*. ACM.
- Rauschenberger, M., Schrepp, M., Cota, M. P., Olschner, S. and Thomaschewski, J. r. (2013) 'Efficient measurement of the user experience of interactive products. How to use the user experience questionnaire (ueq). example: spanish language version.' *IJIMAI*, 2(1) pp. 39-45.
- Reafat, M. M., Ewees, A. A., Eisa, M. M. and Ab Sallam, A. (2012) 'Automated assessment of students arabic free-text answers.' *Int J Cooperative Inform Syst*, 12 pp. 213-222.
- Rensch, C. R. (1992) 'Calculating lexical similarity.' *Windows on bilingualism*, pp. 13-15.

- researchautism. (2016) *Types of Autism*. [Online] [Accessed <http://researchautism.net/autism/types-of-autism>]
- Reynolds-Keefer, L. and Johnson, R. (2011) 'Is a picture is worth a thousand words? Creating effective questionnaires with pictures.' *Practical Assessment, Research & Evaluation*, 16(8) p. 2.
- Roberts, J. and Simpson, K. (2016) 'A review of research into stakeholder perspectives on inclusion of students with autism in mainstream schools.' *International Journal of Inclusive Education*, 20(10) pp. 1084-1096.
- Roberts, K. D. (2010) 'Topic areas to consider when planning transition from high school to postsecondary education for students with autism spectrum disorders.' *Focus on Autism and Other Developmental Disabilities*, 25(3) pp. 158-162.
- Robertson, K., Chamberlain, B. and Kasari, C. (2003) 'General education teachers' relationships with included students with autism.' *Journal of Autism and developmental disorders*, 33(2) pp. 123-130.
- Saadatz, M. N., Pennington, R. C., Welch, K. C. and Graham, J. H. (2018) 'Small-Group Technology-Assisted Instruction: Virtual Teacher and Robot Peer for Individuals with Autism Spectrum Disorder.' *Journal of autism and developmental disorders*, pp. 1-15.
- Saif, A., Ab Aziz, M. J. and Omar, N. (2015) 'Mapping Arabic WordNet synsets to Wikipedia articles using monolingual and bilingual features.' *Natural Language Engineering*, pp. 1-39.
- Schopler, E. and Reichler, R. J. (1972) 'How well do parents understand their own psychotic child?' *Journal of Autism and Childhood Schizophrenia*, 2(4) pp. 387-400.
- Schrepp, M., Olschner, S. and Schubert, U. (2013) 'User Experience Questionnaire (UEQ) Benchmark. Praxiserfahrungen zur Auswertung und Anwendung von UEQ-Erhebungen im Business-Umfeld.' *Tagungsband UP13*,
- Shaalán, K. (2010) 'Rule-based approach in Arabic natural language processing.' *The International Journal on Information and Communication Technologies (IJICT)*, 3(3) pp. 11-19.
- Shaalán, K. F., Attia, M., Pecina, P., Samih, Y. and van Genabith, J. (2012) *Arabic Word Generation and Modelling for Spell Checking*.
- Sharoff, S. (2006) 'Creating general-purpose corpora using automated search engine queries.' *WaCky*, pp. 63-98.
- Shawar, A. and Atwell, E. S. (2004) *An Arabic chatbot giving answers from the Qur'an*. Vol. 2: ATALA.
- Silver, M. and Oakes, P. (2001) 'Evaluation of a New Computer Intervention to Teach People with Autism or Asperger Syndrome to Recognize and Predict Emotions in Others.' *Autism*, 5(3) pp. 299-316.
- Silverbarg, A. and Jönsson, A. (2011) *Subjective and objective evaluation of conversational agents in learning environments for young teenagers*.

- Smith, A. (1998) *Accelerated learning in practice*. A&C Black.
- Snyder, K., Higbee, T. S. and Dayton, E. (2012) 'PRELIMINARY INVESTIGATION OF A VIDEO-BASED STIMULUS PREFERENCE ASSESSMENT.' *Journal of Applied Behavior Analysis*, 45(2) pp. 413-418.
- Stash, N. (2007) Incorporating cognitive/learning styles in a general-purpose adaptive hypermedia system. Vol. 68.
- Tegos, S., Demetriadis, S. and Karakostas, A. (2015) 'Promoting academically productive talk with conversational agent interventions in collaborative learning settings.' *Computers & Education*, 87 pp. 309-325.
- TheNationalAutisticSociety. (2015) *Autism*. London: The National Autistic Society. [Online] [Accessed <http://www.autism.org.uk/about-autism/introduction/what-is-autism.aspx>]
- Tomás, D., González, J. L. V., Bisbal, E. and Moreno, L. (2006) *Experiments with LSA for Passage Re-Ranking in Question Answering*. Citeseer.
- Truong, H. M. (2016) 'Integrating learning styles and adaptive e-learning system: Current developments, problems and opportunities.' *Computers in Human Behavior*, 55 pp. 1185-1193.
- Ultes, S., Schmitt, A. and Minker, W. (2013) *On Quality Ratings for Spoken Dialogue Systems-Experts vs. Users*.
- Waddington, E. M. and Reed, P. (2017) 'Comparison of the effects of mainstream and special school on National Curriculum outcomes in children with autism spectrum disorder: an archive-based analysis.' *Journal of Research in Special Educational Needs*, 17(2) pp. 132-142.
- Wainer, A. L. and Ingersoll, B. R. (2011) 'The use of innovative computer technology for teaching social communication to individuals with autism spectrum disorders.' *Research in Autism Spectrum Disorders*, 5(1) pp. 96-107.
- Walker, M. A., Litman, D. J., Kamm, C. A. and Abella, A. (1997) *PARADISE: A framework for evaluating spoken dialogue agents*. Association for Computational Linguistics.
- Wallace, R. S. (2008) ALICE: Artificial Intelligence Foundation Inc.
- Weizenbaum, J. (1966) 'ELIZA—a computer program for the study of natural language communication between man and machine.' *Communications of the ACM*, 9(1) pp. 36-45.
- Whitaker, P., Joy, H., Harley, J. and Edwards, D. (2001) *Challenging Behaviour and Autism: Making Sense, Making Progress: a Guide to Preventing and Managing Challenging Behaviour for Parents and Teachers*. AAPC Publishing.
- Yahaya, W. A. J. W. and Salam, S. N. A. (2016) 'Smiley Faces: Scales Measurement for Children Assessment.'
- Yahya, W. and Noor, N. (2015) *Decision support system for learning disabilities children in detecting visual-auditory-kinesthetic learning style*.

- Yaman, S. G. (2016) 'Alan Turing: the Unknown Mind.'
- Yin, L., Bickmore, T. and Cortés, D. E. (2010) *The impact of linguistic and cultural congruity on persuasion by conversational agents*. Springer.
- Zhang, J. J., Smith, D. W., Lam, E. T. C., Brimer, J. and Rodriguez, A. (2002) 'Development of an evaluation scale to measure participant perceptions of after-school enrichment programs.' *Measurement in Physical Education and Exercise Science*, 6(3) pp. 167-186.

Appendices

Appendix A – Interview Questions with primary school teachers to develop the tutorial (Arabic version)

درس النظام الشمسي

ماهي أهداف الدرس؟

كيف يتم شرح كل هدف؟

هل يتم استخدام اللوحات التعليمية في الدرس؟

هل يتم استخدام وسائل تكنولوجيا تعليميه ؟ ماهي؟

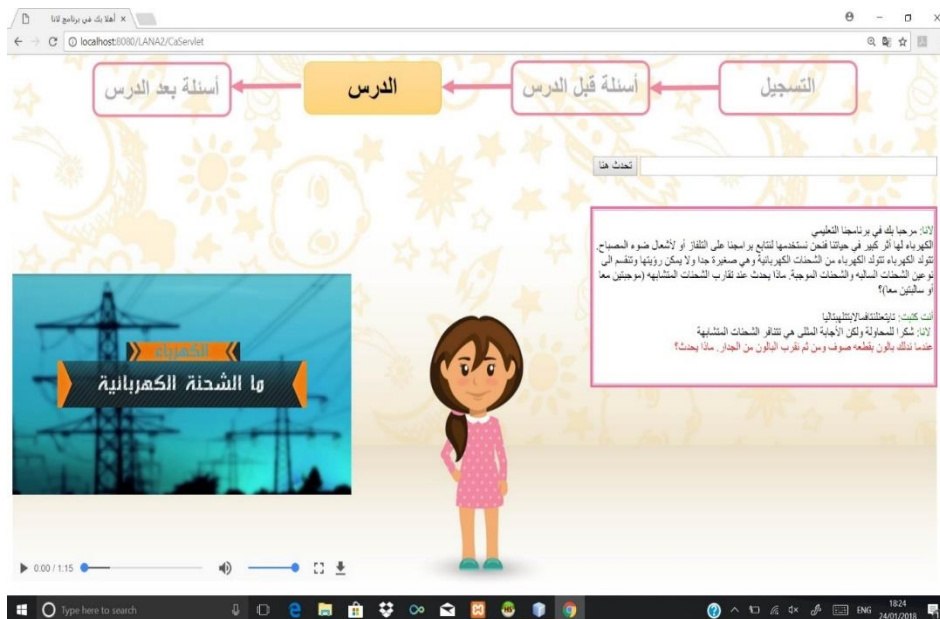
ماهي الأسئلة الافتتاحيه التي يبدأ منها الدرس؟

ماهي الأسئلة التي يتم استخدامها لتقييم فهم الطالب للدرس؟

هل يحبذ الطال الاسئلة المقالية أم الموضوعيه؟

Appendix B – Screenshot of all GUI





Appendix C – Learning style questionnaire (Arabic version)

1. هل تستمتع بالدرس من خلال مناقشة الموضوع مع زميلك ومجموعة من الزملاء
☐ نعم
☐ لا
2. هل انت جيد في عمل أشياء مختلفة كالصلصال والالعاب التركيبات
☐ نعم
☐ لا
3. هل تتذكر وجوه الأشخاص حتى عندما لا تراهم لفترة طويلة
☐ نعم
☐ لا
4. هل تجد من السهولة تذكر الأشياء التي طلبت منك
☐ نعم
☐ لا
5. هل تجد من السهولة تذكر القصص اتي سمعتها
☐ نعم
☐ لا
6. هل تفضل تعلم الاشياء الجديدة من خلال رؤيتها بأكثر من لون وصوره
☐ نعم
☐ لا
7. عندما تتعلم حروف كلمة جديدة هل تفضل رؤية الكلمة جيدا وتتخيل شكلها في مخيلتك
☐ نعم
☐ لا
8. هل تحب استعمال يديك لتشرح لأصدقاءك عن ماتعملة
☐ نعم
☐ لا
9. هل تتذكر حروف الكلمة من خلال التفكير بشكل الكلمة وكتابتها بيدك
☒ نعم
☐ لا

Appendix D – Information sheet for parents

Researcher: Mrs. Sumayh Aljameel **Supervisors:** Dr. James O'Shea – Dr. Keeley Crocket

Address: E113

John Dalton Building
School of computing, Maths and Digital Technology
Chester Street
Manchester, M1 5GD

Phone: +(44)7492464666 **Email:** sumayh.s.aljameel@stu.mmu.ac.uk

You child is being invited to take part in a research study. In order for you to decide whether your child would like to take part, this information sheet explains why the research is being undertaken and what it will involve. For more information, please email Sumayh Aljameel at the address above.

What is the purpose of the study?

The purpose of this research study is to evaluate a new technology for a computer based tutoring system called LANA. The new tutoring system is developed for children on the autistic spectrum age (10-16) to predict their learning styles and provide personalised tutoring within a science domain. The system incorporates graphics, questions/answering and conversation to teach them the science subject.

Why is the study being done?

The study is being undertaken to evaluate the new tutoring system to offer an additional support for learning by children who are on the autistic spectrum.

Why are being asked to allow your child to take part?

As part of this research we are looking for children aged (10 to 16) who have been diagnosed as on the autistic spectrum. The tutoring system has been developed to be appropriate to this age group because they have the ability to write in Arabic language using keyboard.

Does my child have to take part?

No. this is an entirely voluntary study. If you choose not to participate, it will not affect you or your child in any way. If you give your consent for your child to participate, you will be asked to sign a consent form (attached). Even if you give consent, you will still be free to withdraw your child at any time and without giving a reason. Also, your child's class tutor will be able withdraw your child during the study if he or she judges it is right for your child.

Are there any disadvantage or risks in allowing the children to take part?

No, there is no risk or disadvantage beyond ordinary use of a computer in the classroom.

What will happen to your child if you agrees to take part?

Your child will be in a familiar classroom setting with both class teacher and Sumayh Aljameel (the researcher).

Sumayh Aljameel (the researcher) will first introduce the tutoring system to the children.

Your child will be asked first to login to the tutoring system using nickname provided by Sumayh. This prevents personal data from being capture.

Your child will then use the system by having a conversation based tutorial with the system. At the start of the tutorial, a questionnaire will be conducted in order to capture their learning style. Then a tutorial will start on science topic. A pre and pot test will be provided to the child before and after the tutorial in order to evaluate their learning gain. These tests are not intended to measure the children's ability or intelligence and will not be part of their academic record, the overall learning gains of the children involved will be used to measure the performance of LANA.

Who will have access to the data?

Mrs Sumayh Aljameel (the researcher). Sumayh's supervisors will have access to completely anonymous numeric data as part of their supervisory role, but will not have any knowledge of the children involved.

Who is funding the research?

Royal Embassy of Saudi Arabia Cultural Bureau in London

Who they should contact for further information?

Mrs Sumayh Aljameel (the researcher) at the contact address at the top of this sheet.

How will data be anonymised?

The information gathered from pupil is stored securely, coded anonymously, and presented in aggregate form, so that the identity of any individual or home cannot be disclosed outside the research site. The procedures for handling, processing, storage and destruction of data gathered in this study are complaint with Data Protection Act 2008 and the Q.U.B Data Protection Policy. Data will be destroyed within 6 months of either, completion of Sumay's PhD thesis OR completion of the last publication it is used for.

How many participants?

The number of participants will be more than 30 pupils.

Where will the research sessions take place?

The session will take place at your child's home. Every effort will be made to ensure that the research sessions are as enjoyable and relaxed as possible for the children. The total testing time will not exceed 60 minutes.

What is the impact of the study?

At the end of the study the "findings" (scientific principles discovered) will be published and these will contribute to improved education of the community of children on the autistic spectrum and will shape the strategy of future research intended to improved education of the community of children on the autistic spectrum.

Appendix E – Consent Form

Researcher: Mrs. Sumayh Aljameel

Supervisors: Dr. James O'Shea – Dr. Keeley Crocket- Dr. Annabel Latham

Address: E113

John Dalton Building

School of computing, Maths and Digital Technology

Chester Street

Manchester, M1 5GD

Phone: +(44)7492464666

Email: sumayh.s.aljameel@stu.mmu.ac.uk

I have read the information sheet for parents. I understand it and I give my consent for my child _____ (child's name) to take part in the above named research project.

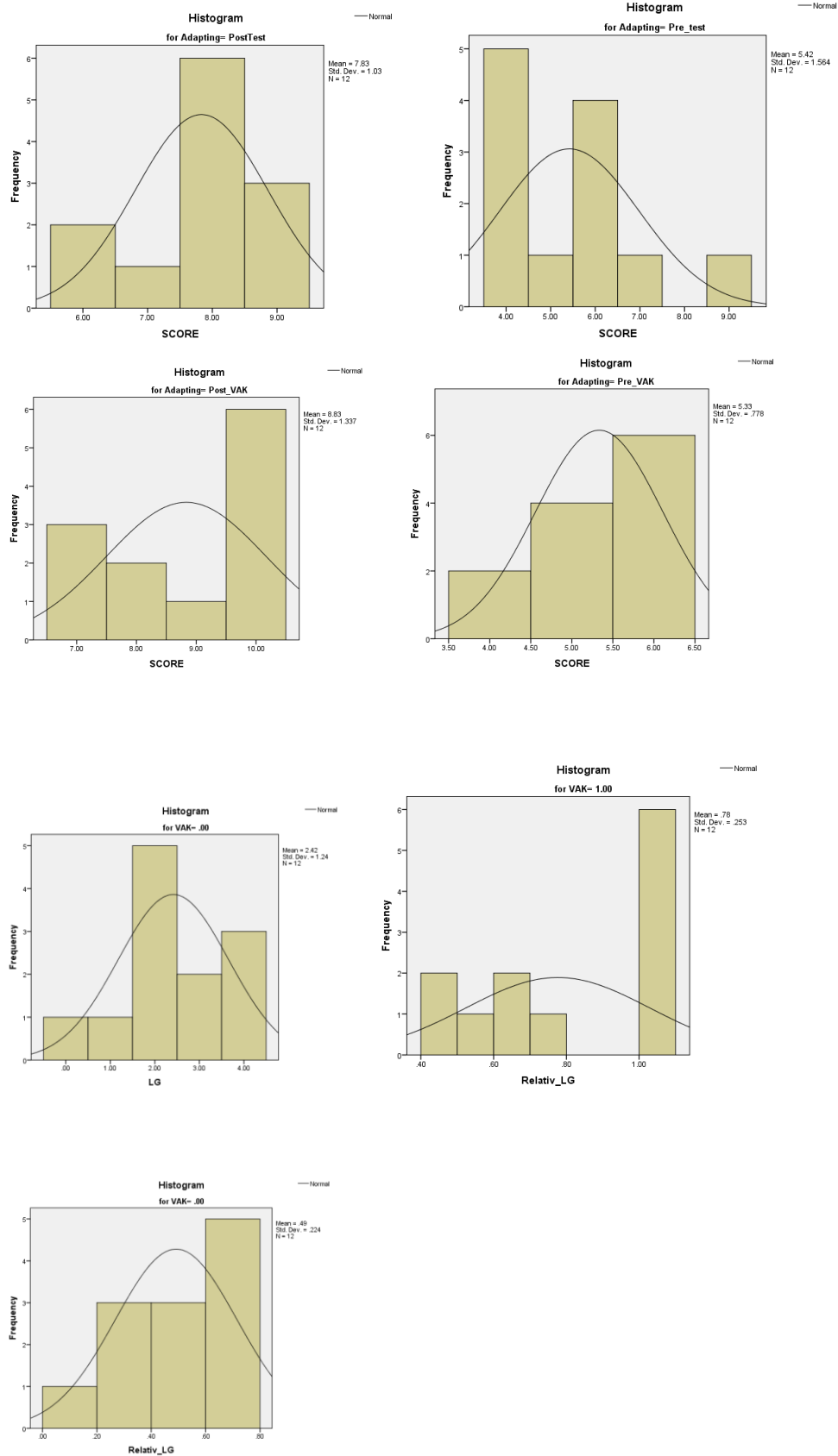
I understand that all data will be kept confidential, it will be destroyed at the end of the study as set out in the information sheet for parents, and that I have the right to withdraw my child from the study at any time.

Print Name: _____

Signature: _____

Date: _____

Appendix F – Normality histogram of the first evaluation



Appendix G – Interview questions with primary school teachers to enhance the tutorial (Arabic version)

من خلال استخدامك لبرنامج لانا 1 . ماهو تقييمك للمعلومات المدرجة به وكيف يمكن تحسينها؟

سيتم ادراج درس الدائرة الكهربائية والمغناطيسية للبرنامج. ماهي أهداف الدرس ؟

كيف يتم شرح كل هدف؟

هل يتم استخدام اللوحات التعليمية في الدرس؟

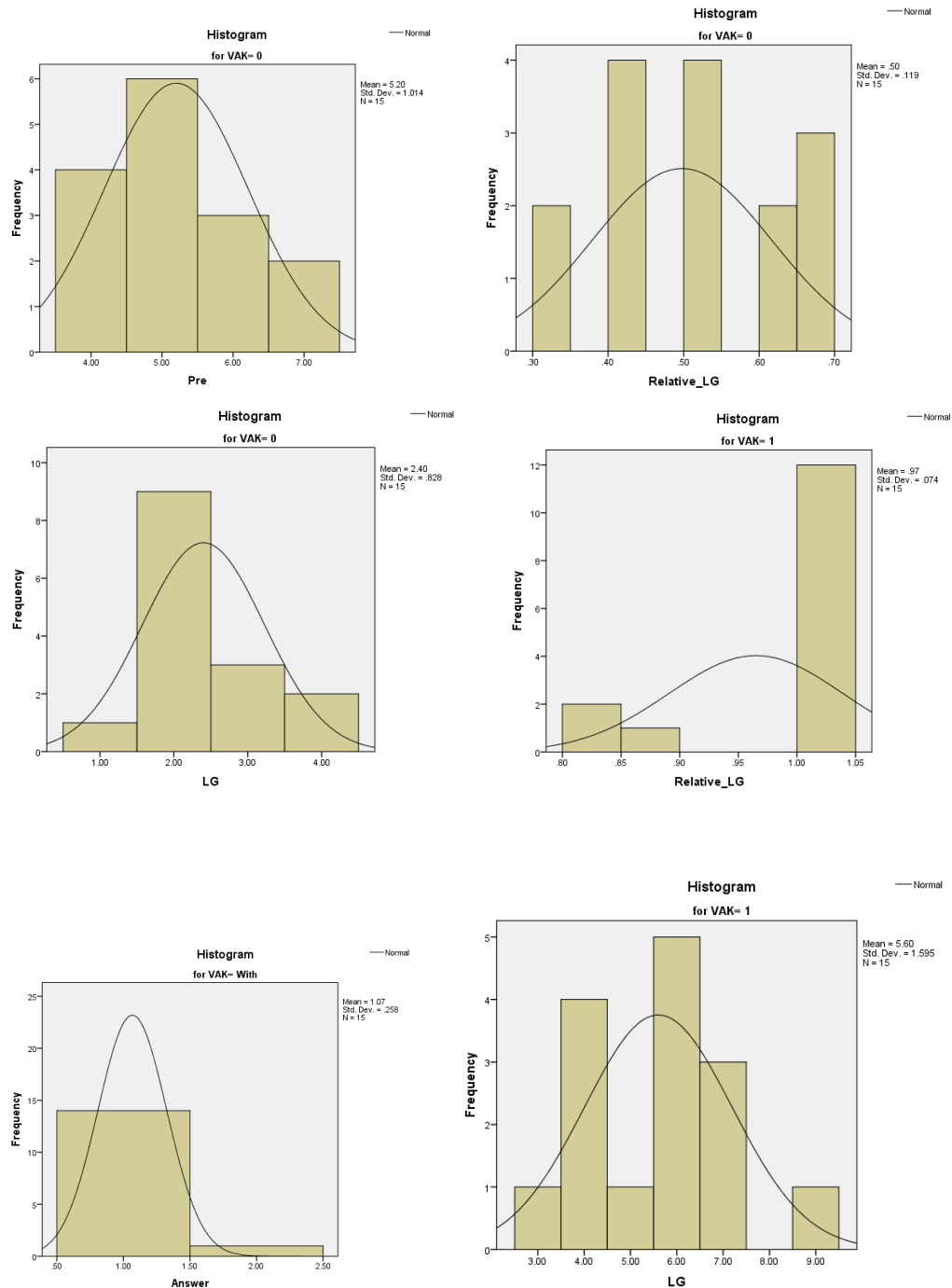
هل يتم استخدام وسائل تكنولوجيا تعليميه ؟ ماهي؟

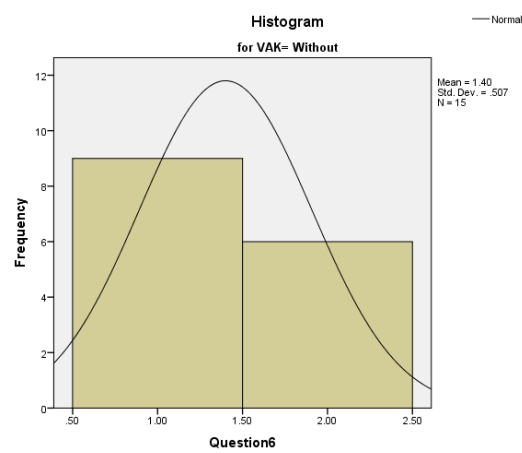
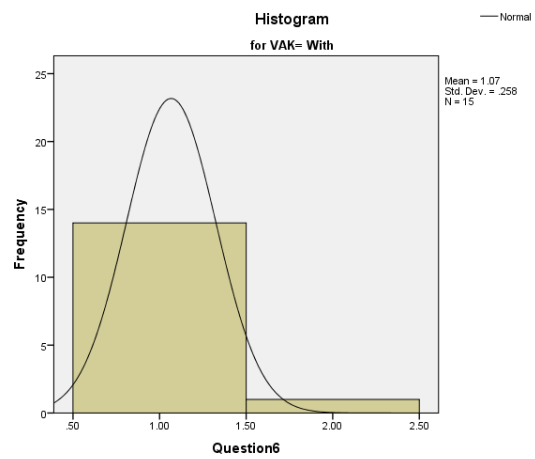
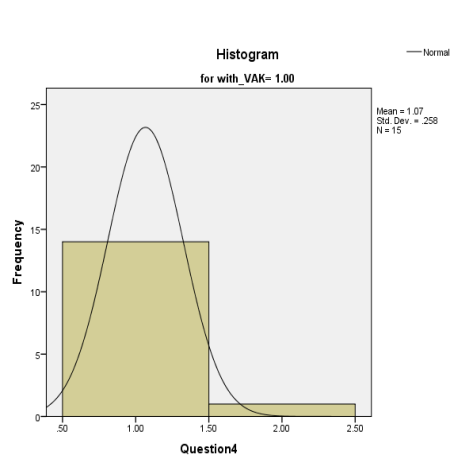
ماهي الأسئلة الافتتاحية التي يبدأ منها الدرس؟

ماهي الأسئلة التي يتم استخدامها لتقييم فهم الطالب للدرس؟

هل يحبذ الطال الاسئلة المقالية أم الموضوعيه؟

Appendix H – Normality Histogram of the second evaluation





Appendix I – Pre-test and Post test questions (Solar system tutorial)

محورها حول لتدور الأرض تستغرق كم تعرف فهل محورها حول وتدور الشمس حول الأرض تدور:

a. 24 ساعة

b. 22 ساعة

c. 12 ساعة

ماذا ينتج عن دوران الأرض حول محورها:

a. الفصول الأربعة

b. الليل والنهار

c. الخسوف والكسوف

تدور الأرض حول الشمس كل سنة. ماهي نتيجة هذا الدوران:

a. الخسوف والكسوف

b. الفصول الأربعة

c. الليل والنهار

كم يستغرق دوران الأرض حول الشمس:

a. 365 يوم

b. 350 يوم

c. 300 يوم

يختلف القمر عن الأرض ب:

a. القمر ليس لديه غلاف جوي ولا ماء

b. القمر ليس لديه صخور

c. القمر أكبر من الأرض

القمر لا يضيئ بنفسه فهو انعكاس لضوء آخر. ما هو:

a. ضوء الشمس

b. ضوء الأرض

c. ضوء الكواكب الأخرى

نرى القمر في أشكال مختلفة خلال الشهر. فما سبب هذا الاختلاف:

a. بسبب دوران الأرض حول الشمس

b. بسبب دوران القمر حول الأرض

c. بسبب دوران الشمس حول القمر

ماهي المدة التي يتم فيها القمر دورته حول الأرض:

a. في 29 يوما أي شهر كامل

b. 20 يوم

c. سنة كاملة

كم عدد أطوار القمر:

a. 6

b. 7

c. 8

خسوف القمر يحدث عندما:

a. تقع الأرض بين القمر والشمس

b. تقع الشمس بين القمر والأرض

c. يقع القمر بين الشمس والأرض

Appendix J – Pre-test and Post-test questions (Tutorial 2: Electrical circuit and the magnetism)

1. ماذا يحدث عند تقارب الشحنات المتشابهة. موجبتين معا أو سالبيتين معا
☐ تتنافر الشحنات
☐ تتجاذب الشحنات
☐ لا يحدث شيء
2. التيار الكهربائي لا يعمل الا في مسار مغلق. ماسم هذا المسار ؟
☐ المجال الكهربائي
☐ الدائرة الكهربائية
☐ الشحنات الكهربائية
3. تتكون الدائرة الكهربائية من ثلاث أجزاء أساسية. وهي المصدر والمقاومة. ماهو الجزء الثالث؟
☐ المغناطيس
☐ الأسلاك
☐ الشحنات
4. كيف نستطيع التحكم بتيار التيار الكهربائي وتشغيل المصباح في الدائرة الكهربائية؟
☐ بقطع الأسلاك
☐ باستخدام المفتاح الكهربائي
☐ باستخدام الشحنات
5. ماذا يحدث عندما ندلك بالون بقطعه صوف ومن ثم نقرب البالون من الجدار؟
☐ يسقط البالون
☐ يلتصق البالون بالجدار
☐ يطير البالون بعيدا
6. ماهي الأجسام الذي يجذبها المغناطيس ؟
☐ المصنوعه من الحديد أو النيكل أو الكوبلت
☐ المصنوعه من الخشب
☐ المصنوعه من الزجاج
7. هناك منطقة محيطة بالمغناطيس تظهر فيها اثار قوته. ماذا تسمى هذه المنطقة؟
☐ القطبين الشمالي والجنوبي
☐ الشحنات الكهربائية
☐ المجال المغناطيسي
8. يتكون المغناطيس من قطبين. ماهما؟
☐ الشمالي والجنوبي
☐ الشرقي والغربي
☐ الأوسط والشمالي
9. ماذا يحدث عندما نقرب قطبين متشابهين لمغناطيسين؟
☐ تتجاذب المغناطيسين
☐ تتنافر المغناطيسين
☐ لا يحدث شيء
10. ماذا يحدث عندما نقرب قطبين مختلفين لمغناطيسين؟
☐ تتجاذب المغناطيسين
☐ تتنافر المغناطيسين
☐ لا يحدث شيء

Appendix K – CBR Check Certificate (DBS)

Enhanced Certificate		Disclosure & Barring Service
Page 1 of 2		
DBS Fee Charged	Certificate Number	001552720048
	Date of Issue:	15 NOVEMBER 2016
Applicant Personal Details		Employment Details
Surname:	ALJAMEEL	Position applied for:
Forename(s):	SUMAYH	CHILD AND ADULT WORKFORCE HEALTHCARE SCIENCE PLACEMENT
Other Names:	NONE DECLARED	Name of Employer:
Date of Birth:	31 JANUARY 1983	MANCHESTER METROPOLITAN UNIVER
Place of Birth:	DAMMAM SAUDI ARABIA	Countersignatory Details
Gender:	FEMALE	Registered Person/Body:
		MANCHESTER METROPOLITAN UNIVERSITY
		Countersignatory:
		MAUREEN AUSTIN
Police Records of Convictions, Cautions, Reprimands and Warnings		
NONE RECORDED		
Information from the list held under Section 142 of the Education Act 2002		
NONE RECORDED		
DBS Children's Barred List information		
NONE RECORDED		
DBS Adults' Barred List information		
NONE RECORDED		
Other relevant information disclosed at the Chief Police Officer(s) discretion		
NONE RECORDED		
Enhanced Certificate		
This document is an Enhanced Criminal Record Certificate within the meaning of sections 113B and 116 of the Police Act 1997.		
THIS CERTIFICATE IS NOT EVIDENCE OF IDENTITY		Continued on page 2
Disclosure and Barring Service, PO Box 165, Liverpool, L69 3JD Helpline: 03000 200 190		

Appendix L – Manchester Metropolitan Ethic Approval

FACULTY OF SCIENCE AND ENGINEERING



MEMORANDUM

TO Sumayh Aljameel
FROM Francesca Robinson
DATE 17th March 2016
SUBJECT Application for Ethical Approval (SE151615)

On the 17th March 2016 the Head of Ethics for Science & Engineering considered your application for Ethical Approval (SE151615) entitled "*An Arabic Conversational Intelligent Tutoring System for Autism*". The application has been granted Favourable Opinion and you may now commence the project.

MMU requires that you report any Adverse Event during this study immediately to the Head of Ethics (Dr Nick Costen) and the Administrator (Megan Schofield). Adverse Events are adverse reactions to any modality, drug or dietary supplement administered to subjects or any trauma resulting from procedures in the protocol of a study.

An Adverse Event may also be accidental loss of data or loss of sample, particularly human tissue. Loss of human tissue or cells must also be reported to the designated individual for the Human Tissue Authority licence. Please notify Professor Craig Banks of any issues relating to this.

If you make any changes to the approved protocol these must be approved by the Faculty Head of Ethics. If amendments are required you should complete the MMU Request for Amendment form (found on the Graduate School website) and submit it to the Administrator.

Regards

Francesca Robinson
Research Administrator
All Saints North

Appendix M – Author Publications

A Review of Current Technology-Based Intervention for School Aged Children with Autism Spectrum Disorder

Sumayh S. Aljameel^(✉), James D. O'Shea, Keeley A. Crockett,
and Annabel Latham

Department of Computing, Math and Digital Technology,
Manchester Metropolitan University, Manchester, England, UK
Sumayh.s.aljameel@stu.mmu.ac.uk,
{j.d.oshea,k.crockett,a.latham}@mmu.ac.uk

Abstract. For individuals with autism spectrum disorder (ASD), the use of Computer technology to provide intervention in learning is promising. This review focuses on research that has used technology to improve the performance for school aged (10–16) children with ASD. This paper reviews technologies that enhanced intervention, which target three cognitive domains: (1) languages and literacy, (2) social skills, and (3) emotion recognition. A review of the literature from 2005 to the end of 2015 identified 19 studies that documented efficacy in order to determine whether empirical findings support technology as an evidence-based practice. The conclusion reports that it is important to support development, evaluation, and clinical usage of technology-based intervention for individuals with autism spectrum disorders. Future directions for research and practice with each technology are discussed.

Keywords: Autism spectrum disorder · Computer · Technologies · Intervention

1 Introduction

Autism spectrum disorder (ASD) is a lifelong developmental disability, which affects communication skills, social skills and repetitive behavior [1]. ASD has four main subtypes of autism: Autistic Disorder, Asperger Syndrome, Pervasive Developmental Disorder and Childhood Disintegrative Disorder. Autism is a spectrum condition that means people with autism have shared specific difficulties, but they will be affected by their condition in different ways. Some autistic people have the ability to live relatively independently where others need a specialist support to live their lives. People with autism may also have sensitivity to sounds, lights, colours, touch, smells, and tastes. Autism affects three areas: communication (verbal and non-verbal language), social interaction (understanding and recognizing emotions and expression), and repetitive behaviours (adaptation to the environment). Over the past two decades, the cases of autism have increased worldwide. It is reported that more than one out of every 100 people has autism in UK [2]. Some studies suggests that ASD is genetic [3, 4].

© Springer International Publishing AG 2018

Y. Bi et al. (eds.), *Proceedings of SAI Intelligent Systems Conference (IntelliSys) 2016*,
Lecture Notes in Networks and Systems 16, DOI 10.1007/978-3-319-56991-8_63

The genetic cause can be identified in 20%–25% of children with ASD using standard medical genetic evaluation techniques [5]. However, observations of behavioural aspect are the only way to diagnose ASD. The average age of diagnosis of ASD in UK is over three years old, although another study recommends that all children should be screened for ASD earlier at 18 to 24 months [6].

Several studies reported that people with ASD have the ability to deal with technology and computers better than human [7–9]. Computer Applications offer a structured environment that is preferred by people with autism because of routine and repetitive behaviours [10]. Virtual tutors have shown to benefit children with ASD, who have difficulties in social interactions [11]. Traditional teaching using human tutors is a challenge for children with autism whereas virtual tutors can meet the individual child needs. This is because some autistic children are uncomfortable looking at faces; the virtual tutor face looks like a cartoon, and when the child becomes familiar with it, the complexity of expression can be increased. In the same way, the session content can be adjusted depending on the child's current level of need and interest. There is some evidence to suggest that autistic children those taught by virtual tutors gain more information than who were taught by human tutors. [12]. Virtual tutors have many benefits, for example, children learn and practice their skills independently with no pressure on the person who works with them. Consequently, the human tutor will focus on complex aspects and tasks while the virtual tutor will focus on the general topics. [13]. Another advantage from using the virtual tutor is that learners can practice at any time of the day because the virtual tutor never gets impatient or tired, which suits with autistic people who have abnormal sleep patterns. [14]. Regarding developing nonverbal skills, it is reported that virtual tutors can be useful in modelling the learner's behaviours, such as gaze and facial expression [15].

The main purpose of this paper is to:

- review different technologies that are used to enhance the emotion recognition, social skills and language skills for children with ASD aged (10 to 16) or covered this age.
- review different contexts and evaluation methods that are used for measuring the efficiency of technology-based interventions for children with ASD in the age group of 10 to 16 years.
- Recommend the future directions for research and practice with each technology.

This paper is organized as follows: Sect. 2 describes the methods that are used in this review. Section 3 includes the technology-based interventions in the three cognitive domains: (1) languages and literacy, (2) social skills, and (3) emotion recognition. Section 4 contains the conclusion.

2 Methods

2.1 Literature Search Procedures

Online searches were conducted within multiple research databases, such as Google Scholar and PubMed using the following search terms: 'Virtual agent and autism',

‘autism and technology, ‘computer-based intervention and autism’ and ‘robotics and autism’. In addition, relevant articles were involved in this review (e.g. Wass and Porayska-Pomsta [16], McCleery [17]).

2.2 Inclusion Criteria

Our inclusion criteria for this review were as follows:

- The research tested their system using participants in the age group of 10 to 16 or covered this age. The reason for choosing this age group is that many researchers concentrate on the younger children (3–7 years old or pre-school age).
- The system is to enhance one or more social skills, emotion recognition, and language and literacy.

A total of 19 articles from 2005 to the end of 2015 met the inclusion criteria as technology interventions for children with ASD. Table 1 contains information from the

Table 1. Summary of the studies featured in this review

Author	year	Title	Participants number (N) and Age	Diagnosis criteria	Specific Target	Context	Evaluation methods	Description of training	Results
Alves et al. [36]	2013	LIFEinGAME: Prototype: A Serious Game about Emotions for Children with Autism Spectrum Disorders	N: 11 Age: 5-15	ASD	Emotion recognition	Matching the facial expression faces.	Video recorded each child during an open 15 minutes game session. Footages were qualitatively analyzed by two independent researchers regarding motivation to play and game usability.	A serious game that proposes to enhance facial and emotional recognition skills in children with Autism Spectrum Disorders (ASD).	Improvements within the training.
Baumin ger-Zivety et al. [36]	2013	Increasing social engagement in high-functioning children with ASD using collaborative technologies in the school	N: 22 Age: Mean 9.8 (standard deviation = 10.7)	ASD	Social skills	A series of social vignettes, the children who were able to learn about collaboration by selecting and discussing alternate social solutions.	Assessment in the socio-cognitive area included concept perception measures, problem solving, Theory of Mind (ToM), and a dyadic drawing collaborative task to examine change in children's social engagement.	Two computer programs (one tabletop and one laptop) designed to teach social conversation based around principles of Cognitive Behavioral Therapy.	Improvements on concept clarification and problem-solving, some improvement on ToM.
Bernardini et al. [27]	2014	ECHOES: An intelligent serious game for fostering social communication in children with autism	N: 29 Age: 4 to 14	ASD	social communication skills	12 activities in a sensory garden populated by Andy and interactive “magic” objects	pre and post table-top assessments	A serious game for supporting social communication skills of children with ASD.	Improvements within the training.
Cheng et al. [28]	2015	Using a 3D Immersive Virtual Environment System to Enhance Social Understanding and Social Skills for Children With Autism Spectrum Disorders	N: 3 Age: 10-13	ASD	social communication skills	virtual bus stop and classroom environment	score for appropriate answers	A three-dimensional social understanding system with a head-mounted display to improve social understanding and skills in three children with ASD.	The findings indicate that participants' targeted behaviors improved from baseline to intervention through maintenance following their use of the system.
Chien et al. [29]	2015	ICAN: A tablet-based pedagogical system for improving communication skills of children with autism	N: 11 Age: 5-16	ASD	social communication skills	digital picture cards contains words	Pre and post interview questions	This paper presents ICAN, a tablet-based system that adapts the successful aspects of the traditional the Picture Exchange Communication System (PECS)	ICAN reduced content preparation time by over 70% while also reducing children with autism's willingness to learn and interact with others.
Chen et al. [37]	2015	Augmented reality-based self-facial modeling to promote the emotional expression and social skills of adolescents with autism spectrum disorders	N: 3 Age: 10-13	ASD	emotional expression and social skills	six facial expressions that communicated basic emotions (happiness, sadness, fear, disgust, surprise, and anger) associated with a short story	Questions during the story and assess the correct answers	The AR system provided three-dimensional (3-D) animations of six basic facial expressions overlaid on participants faces to facilitate practicing emotional judgments and social skills.	Improve the appropriate recognition and responses to facial emotional expressions seen in the situational task.
Colonna-Martin et al. [22]	2005	Using computer-aided instruction and the novel word reading approach to teach word identification.	N: 1 Age: 12	ASD	Language and literacy	word identification	Fifteen words were randomly selected for each student from the pool of words that were identified with 33% accuracy or less.	Power Point software. Each slide had a visual and auditory component. The first slide for each word began by showing the student the entire word and encouraging her to slowly say the word aloud as the word was slowly pronounced by the computer program.	Student made accurate target words in all conditions

Table 1. (Continued)

Escobedo et al. [35]	2012	MOSOCO: A Mobile Assistive Tool to Support Children with Autism Practicing Social Skills in Real-Life Situations	N: 3 Age: 8-11	ASD	Social skills	Eye contact, space and proximity, start an interaction, asking questions, sharing interest, and finish an interaction.	Interview and videos recording are used. A quantitative analysis was conducted to estimate the time students spent in different interactions.	MOSOCO facilitates practicing social skills in real-life situations, guiding students through the six fundamental social skills of the Social Compass curriculum.	Increases both quantity and quality of social interactions, reduces social and behavioral missteps, and enables the integration of children with autism in social groups of neurotypical children.
Grynszpan et al. [36]	2008	Multimedia interfaces for users with high-functioning autism: an empirical investigation	N: 10 Age: 12	HFA	Social Skills	Game, called "What to choose?" (Fig. 2), prompted the user to choose one of three assertions about the dialogue.	"pre-post test" design to assess a 3-month training.	Text-based software that trained subjects to understand dialogues that contain pragmatic subtleties, sometimes with a static computer-generated face and sometimes not.	Improvement with the no-face condition
Hietroni and Shalom [37]	2005	From Logos to Orthographic Symbols: A Multilevel Fading Computer Program for Teaching Neoverbal Children With Autism	N: 6 Age: 10-13	ASD	Language and literacy	teaching orthographic symbols	the preliminary generalization test was repeated during regular activities in the classroom.	Matching the right words to the sample	All children obtained and most generalized to actual food items
Masera and Bender [38]	2006	Read my lips: The importance of the face in a computer-animatd tutor for vocabulary learning by children with autism.	N: 3 Age: 8-12	ASD	Language and literacy	vocabulary lessons involving the association of pictures and spoken words.	pre-test, presentation, recognition, elicitation, and post-test	Animated tutor (Bibbi) with and without a face to identify pictures and verbal production of vocabulary words.	Retention was good and effective with the face condition more than no face condition
Jonman et al. [39]	2008	Effectiveness of virtual reality for teaching cross-crossing skills to children and adolescents with autism	N: 6 Age: 8-16	ASD	Social skills	cross a street safely via a desktop screen, crossing virtual environment	A preference safety checklist scored from taped video clips	VR exercises on road safety	Improvements within the training
Knight et al. [34]	2015	An Exploratory Study Using Science eTexts With Students With Autism Spectrum Disorder	N: 4 Age: 10-12	ASD	Language and literacy	science e-books	Number of correct responses	Supported electronic text (eText), or text altered to provide support, to promote comprehension of science content for students with disabilities.	Three of four students increased mean number of correct responses
LaCava et al. [38]	2007	Using assistive technology to teach emotion recognition to students with Asperger	N: 8 Age: 8-11	AS	Emotion recognition	emotion recognition	Emotion recognition testing was conducted using a computer as pre- and post-intervention.	Silent movie clips, voice clips, written descriptions (separate modalities)	Improvement on all measures administered
Lahiri et al. [32]	2013	Design of a Virtual Reality-Based Adaptive Response Technology for Children With Autism	N: 8 Age: 13-18	HFA	social communication skills	24 social tasks in which avatars narrated personal stories to the participants with questions at the end.	captures eye data of a participant interacting with an avatar using eye-tracker goggles	VR-based interactive system with gaze-sensitive adaptive response technology that can seamlessly integrate VR-based tasks with eye-tracking techniques to intelligently facilitate engagement in tasks relevant to advancing social communication skills.	quantitative improvement of performance metric increased viewing of the face of the communicator
Milne et al. [33]	2011	Designing and evaluating interactive agents as social skills tutors for children with autism	N: 14 Age: 6-15	3 HFA 10 AS 1 ASD	social communication skills	How to deal with bullying	survey	Autonomous agents as social skills tutors for teaching children with ASD conversation skills and how to deal with bullying.	Improvements within the training
Moore et al. [39]	2005	Collaborative virtual environment technology for people with autism	N: 14 Age: 8-16	ASD	Emotion recognition	avatar representations for four emotions: happy, sad, angry, and frightened.	observed responses of the participants to the questions against the responses that would be expected were they to be selected by chance.	Avatar (NB only 4 emotions: happy, sad, angry and frightened)	Improvements within the training.
Rice et al. [34]	2015	Computer-Assisted Face Processing Instruction Improves Emotion Recognition, Mentalizing, and Social Skills in Students with ASD	N: 31 Age: 5-11	ASD	social skills and facial recognition	objects, numbers, or faces	Pre- to post-intervention mean score on mentalizing assessments as compared to participants in the control condition.	Computer-based social skills intervention called FaceSayTM was associated with improvements in affect recognition, mentalizing, and social skills of school-aged children with Autism Spectrum Disorder (ASD).	Improvement in their affect recognition and mentalizing skills, as well as their social skills.
Tanaka et al. [40]	2010	Using computerized games to teach face recognition skills to children with autism spectrum disorder: the Let's Face It! program	N: 42 Age: 10-5	ASD	Emotion recognition	face and object	the independent variables were group (active treatment, waitlist control) and time point (Time 1, Time 2), and the dependent variable was total percentage accuracy on the given subset.	Interactive computer games targeting face recognition skills played at home	Improvements within the training

ASD: autism spectrum disorder; AS: Asperger's syndrome; HFA: high-functioning autism; LFA: low-functioning autism; PDD-NOS: pervasive developmental disorder-not otherwise specified

studies (Author, Title, Participants number (N) and Age, Diagnose criteria, Specific Target, Context, Evaluation methods, Description of training, Results).

3 Technology-Based Intervention

Technology-based interventions are the most studied for children with ASD. Technology has been used to teach a variety of skills, such as how to recognize and predict emotions [18], enhance problem solving [19], improve vocabulary [20], and improve reading and communication skills [21]. Our review focuses on technologies that enhanced intervention, which targets three cognitive domains: (1) languages and literacy, (2) social skills, and (3) emotion recognition.

3.1 Languages and Literacy

Language learning is often delayed in children with ASD [20]. Therefore, many interventions focus on enhancing language and communication by developing programs that are used to teach ASD children language content. Four articles are summarized in this section that used technology to enhance the language and literacy skills.

The first one is developed by Coleman-Martin et al. [22], which is Power Point software. Each slide has two components: visual and auditory. Each word has a slide that showing the entire word and encouraging the child to slowly say the word aloud as the word was slowly pronounced by the computer program. The experiment is conducted using one student with ASD who's age is 12 and two others with different disabilities. The result showed an improvement within the training. However, the software used in this experiment is (PowerPoint), which is considered as a low technology intervention because it has limited properties such as the dynamic feedback. Furthermore, one child with ASD was involved in this research, which is considered as a disadvantage of this research because the sample size is assessing the utility of interventions. This allows for the long-term refinement and development of training targets. Hetzroni and Shalem [23] develop an application, which is a computer program for teaching nonverbal children with autism. The program asked the children to match the sight words to the sample. Six children aged 10 to 13 were involved in this research and the result shows that all children matched the right answer and most of them generalized to actual food items. However, sample sizes are small and no control group was included in this research. Massaro and Bosseler [20] developed a computerized tutor (Baladi) to teach vocabulary and functional language use to children with ASD with and without a face to identify pictures and verbal production of vocabulary words. They evaluated whether the face would increase the rate of learning for receptive measures. Five children aged 8 to 13 with ASD are involved in this research and the Retention was good and effective with the face condition more than no face condition but they did not perform statistical tests on their small sample.

The last study in this review related to enhancing the language skills is Knight et al. [24]. They developed (eText), or text altered to provide support, to promote comprehension of science content. Four participants aged 10 to 12 years were involved in this

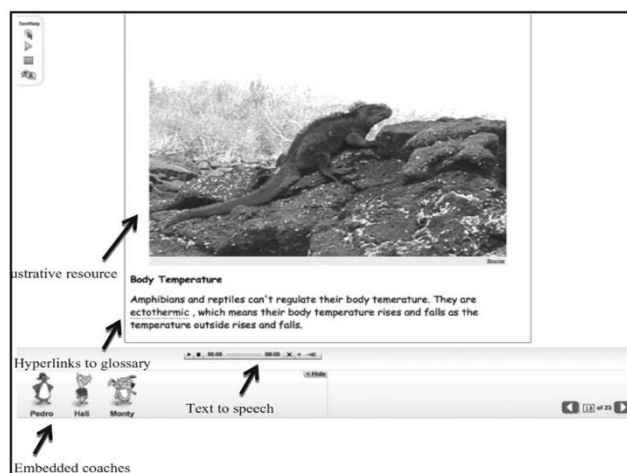


Fig. 1. Screenshot of supports offered in eText system [24]

research, and the result showed that three of four students increased the mean number of correct responses. However, the sample size was small and there is no control group involved in this study (Fig. 1).

3.2 Social Skills

Individuals with ASD have a poor social understanding and social awareness [25]. This causes difficulty in building a relationship or integrating into a mainstream classroom and other environments [19]. Thus, it is important to develop an effective social skills interventions, which targeting individuals with ASD. This section will outline ten studies of programs that have used innovative technology to teach social skills. Bauminger-Zviely et al. [26] designed two computer programs to teach social conversation. The programs were developed based on principles of Cognitive Behavioural Therapy (CBT). 22 children with ASD were involved in the study and the results showed improvements on the concept of the problem-solving, and on Theory of Mind (ToM). However, the study does not contain a control group. Moreover, the results should be treated with caution because of the differences in the children's age (mean = 9.8 years, standard deviation = 10.7 years). Bernardini et al. [27] developed an autonomous virtual agent (ECHOES), which is a serious game for supporting social communication skills of children with ASD. A total of 29 children aged 4 to 14 years were involved in this study. The results indicated significant improvements for the social behaviours using gaze and for children's speech to the agent. However, the system does not have a strict goal-directed training because it focuses on free play and exploration. In addition, children of different ages (4–14) were involved in the experiment, which made it difficult to create an environment suitable for all age groups (Fig. 2).



Fig. 2. A child using the ECHOES software, taken from Bernardini et al. [27]

Cheng et al. [28] report on developing a three-dimensional system that teaches social skills with a head-mounted display. Three children with ASD were involved in the study in the age group of 10–13 years. The results reported that participants' targeted behaviors improved from baseline to intervention through maintenance following their use of the system. The main limitation was the small sample size, which may have limited the generalizability of the data. Furthermore, the study used children with ASD who had above average IQs and selective attention, which may have helped them, learn quickly.

Chien et al. [29] report on developing iCAN, a tablet-based system that adopts the successful aspects of the traditional Picture Exchange Communication System (PECS). Eleven children with ASD were involved in the study age 5–16 years old. It is reported that iCAN reduced content-preparation time by over 70% because the system allows user to Add and edit picture cards and categories by using Card Creator and Editor mode. After selecting the image from any source (snapshot, hand-drawing, or any existing digital pictures), the user can then record the associated word's pronunciation and customize the representative picture of the categories based on their preference. The system also enhanced the willingness of children with ASD to learn and interact with others. This study used a control group to recognise the efficiency of using iCAN. Therefore, the system successfully conducted quantitative analyses on interview questions in order to compare the time that was spent in producing and retrieving the picture cards between users who used traditional Picture Exchange Communication System (PECS) and users who used iCAN.

Escobedo et al. [25] developed MOSOCO for teaching the social skills in real-life situations. Students are guided through six fundamental social skills of the Social Compass curriculum. Three children with ASD were involved in the study in the age group of 8 to 11 years. The results indicated that the system increases the social interactions and reduces social and behavioural missteps. Moreover, the results showed that the system enables the integration of autistic children in social groups of neurotypical children. Although control group is used, the small number of participants limited this study.

Grynszpan et al. [30] report on developing Text-based software, which teaches subjects to understand dialogues that contain pragmatic subtleties, such as, irony, metaphors or sarcasm. The study sometimes used a static computer-generated face and sometimes not. Ten children with High Functional Autism (HFA) were involved in the study, and the result indicated an improvement on recognize the facial expressions with the no-face condition.

Josman et al. [31] report on developing virtual reality for teaching street-crossing skills to children and adolescents with autism aged 8 to 16 years old and the results showed improvements in the ability to cross the virtual street safely within the training. The limitation here is that the research did not apply the same experiment in the real life.

Lahiri et al. [32] developed a virtual reality (VR)-based interactive system with Gaze-sensitive adaptive response technology. The system integrates VR-based tasks with eye-tracking techniques to teach social communication skills. A total of 8 HFA children were involved in the study, and the results reported a quantitative improvement of performance metric increased viewing of the face of the communicator. However, this study has some disadvantages. For example, the wearable eye-tracker and the bi-directional conversational module required use of a mouse, which may not suitable for children with low functioning ASD. Furthermore, the sample size is small and this may affect the results.

Milne et al. [33] report on developing autonomous agent as social skills tutor. It is used to teach autistic children conversation skills and how to deal with bullying. The study involved three children with HFA, ten children with Asperger Syndrome (AS), and one child with ASD. Their age was from 6 to 15 years and the results indicate improvements within the training.

Rice et al. [34] developed computer-based social skills intervention called Face-SayTM, which was used to enhance the memorizing, affect recognition, and social skills of school-aged children with ASD. 31 autistic children were involved in the study in the age group of 5 to 11 years. The results report an improvement in their memorizing skills, affect recognition, and their social skills.

3.3 Emotion Recognition

People with ASD usually have difficulty in understanding and identifying others' emotions [35]. One way to develop this skill is to train individuals in emotion identification and recognition. Five articles are reviewed in this section, which is related to enhancing emotion recognition using technologies.

Alves et al. [36] developed a serious game that enhances emotional and facial recognition skills in autistic children. Eleven children with ASD are involved in the study aged 5 to 15 years. LIFE is GAME contains five games modes. The first mode is Recon Mee-Free where the player should join a thought, such as, join a gift, to the character's facial expression. The second mode is Recon Mee-Match where the player needs to match character's facial expressions to other facial expressions. The third mode is Sketch Mee where the player uses the drawn facial expressions and creates a video to watch the changing of the character's face. The fourth mode is Memory Game

where the player should join facial expressions of models. The fifth mode is Build the Face where the player needs to draw a facial on the character's face based on a target expression showed in a card. The results show an improvement in the children performance within the training (Fig. 3).



Fig. 3. Game mode “Build the Face” from Alves et al. [36]

Chen et al. [37] developed the augmented reality-based (AR) system to enhance emotional judgments and social skills by providing three-dimensional (3-D) animations of six basic facial expressions overlaid on participant's faces. Three participants aged 10 to 13 years were involved in this study. The results indicated that AR intervention could improve the appropriate recognition and response to facial emotional expressions seen in the situational task. However, the small number of participants and no control group limited the results. LaCava et al. [38] developed an assistive system to train emotion recognition to students with AS. They used Silent movie clips, voice clips, and written descriptions (separate modalities). Eight children with AS were involved in this study between the ages of 8 and 11 years. The results showed an improvement on face and voice emotion recognition for basic and complex emotions that were in the software. Again the sample size was small and no control group was used.

Moore et al. [39] developed an Avatar to recognize four emotions: sad, happy, frightened, and angry. The research question here is to determine if children with ASD could understand basic emotions as represented by a humanoid avatar. A total of 34 children with ASD were involved in the study, and the results indicated an improvement within the training. Over 90% of the participants accurately recognized emotions displayed by avatar representations. Although the sample size was not small comparing with other studies, the study did not use control group to emphasise of the positive results.

Tanaka et al. [40] administered an interactive gameto enhance face recognition skills, which can be played at home. Their research question was whether face recognition skills can be enhanced through a direct training intervention. 42 children were involved in the study with average age 10.5. The results show that there was an efficient improvement in their analytic recognition of mouth features and holistic recognition of a face based on its eyes features within the training. However, it is

reported that these improvements failed to generalise to emotion and face recognition tasks that use various emotions and faces to the participants involved in the training set.

4 Conclusion

The review contributes to the literature in three important ways. First, it focused the review on technologies that are used as an intervention for children with ASD aged 10–16 years old. Second, this review concentrated on technologies that enhanced intervention, which targets three cognitive domains: (1) languages and literacy, (2) social skills, and (3) emotion recognition. Third, this review was comprehensive in that it provided information across the participant's number, diagnoses criteria, specific target, context, evaluation methods, description of training, and results.

Nineteen articles were reviewed in this paper, and the results indicated that there is a sufficient improvement in learning for children with ASD within the computerised training paradigm. However, many of the existing technologies have limited capabilities in their performance and thus limit the success in the clinical approach of children with ASD:

- Due to difficulties in recruiting participants, the sample size was small in all studies, and most have no control group. Future research should increase the participants number and to involve a control group without Cognitive Behavioural Treatment (CBT) and a control group with CBT and without collaborative technology.
- Participants included in this review were diagnosed with ASD, and have differences in functional abilities [21]. Future research should examine the effect of participant characteristics on program results. For example, a program that teaches children with ASD advanced language may not be appropriate for children with ASD who uses one to two words to communicate.
- Most studies did not assess whether the participants demonstrate the target behaviors in daily life or not. All the participants performed their tasks in the system with improved self-efficacy after the intervention.

In conclusion, technology-based intervention for children with ASD is a promising strategy for treating them. However, future research must demonstrate the effectiveness of such technologies by using larger number of participants in order for the technologies be used by individuals, schools and mental health agencies in real-life scenarios. A meta-analysis review is recommended as a future work, which will provide a quantitative estimate for the effect of a technology intervention for children with ASD.

References

1. American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders. American Psychiatric Publishing, Arlington (2013)
2. The National Autistic Society: Autism (2015). <http://www.autism.org.uk/about-autism/introduction/what-is-autism.aspx>

3. Miles, J.H.: Autism spectrum disorders-a genetics review. *Genet. Med.* **13**, 278–294 (2011)
4. Wiggins, L.D., Levy, S.E., Daniels, J., Schieve, L., Croen, L.A., DiGuseppi, C., et al.: Autism spectrum disorder symptoms among children enrolled in the study to explore early development (SEED). *J. Autism Dev. Disord.* **45**, 3183–3194 (2015)
5. Carter, M.T., Scherer, S.W.: Autism spectrum disorder in the genetics clinic: a review. *Clin. Genet.* **83**, 399–407 (2013)
6. Zablotzky, B., Bramlett, M., Blumberg, S.J.: Factors associated with parental ratings of condition severity for children with autism spectrum disorder. *Disabil. Health J.* **8**, 626–634 (2015)
7. Banire, B., Jomhari, N., Ahmad, R.: Visual Hybrid Development Learning System (VHDLs) framework for children with autism. *J. Autism Dev. Disord.* **45**, 3069–3084 (2015)
8. Tartaro, A., Cassell, J., Ratz, C., Lira, J., Nanclares-Nogués, V.: Accessing peer social interaction: using authorable virtual peer technology as a component of a group social skills intervention program. *ACM Trans. Accessible Comput.* **6**, 1–29 (2014)
9. Picard, R.W.: Future affective technology for autism and emotion communication. *Philos. Trans. Royal Soc. B Biol. Sci.* **364**, 3575–3584 (2009)
10. Murray, D.: *Autism and Information Technology: Therapy With Computers*. Brookes, Baltimore (1997)
11. Parsons, S., Cobb, S.: State-of-the-art of virtual reality technologies for children on the autism spectrum. *Eur. J. Spec. Needs Educ.* **26**, 355–366 (2011)
12. Grynspan, O., Martin, J.-C., Nadel, J.: Multimedia interfaces for users with high functioning autism: an empirical investigation. *Int. J. Hum. Comput. Stud.* **66**, 628–639 (2008)
13. Massaro, D.: Symbiotic value of an embodied agent in language learning. Presented at the Proceedings of the 37th Hawaii International Conference on System Sciences (2004)
14. Limoges, E., Mottron, L., Bolduc, C.B.: Claude Godbout, Roger, atypical sleep architecture and the autism phenotype. *Brain* **128**(5), 1049–1061 (2005)
15. Marcus, A., Wilder, D.: A comparison of peer video modelling and self video modelling to teach textual responses in children with autism. *J. Appl. Behav. Anal.* **42**(2), 335–341 (2009)
16. Wass, S.V., Porayska-Pomsta, K.: The uses of cognitive training technologies in the treatment of autism spectrum disorders. *Autism* **18**, 851–871 (2014)
17. McCleery, J.P.: Comment on technology-based intervention research for individuals on the autism spectrum. *J. Autism Dev. Disord.* **45**, 3832–3835 (2015)
18. Silver, M., Oakes, P.: Evaluation of a new computer intervention to teach people with autism or asperger syndrome to recognize and predict emotions in others. *Autism* **5**, 299–316 (2001)
19. Bernard-Opitz, V., Sriram, N., Nakhoda-Sapuan, S.: Enhancing social problem solving in children with autism and normal children through computer-assisted instruction. *J. Autism Dev. Disord.* **31**, 377 (2001)
20. Massaro, D.W., Bosseler, A.: Read my lips: the importance of the face in a computer-animated tutor for vocabulary learning by children with autism. *Autism* **10**, 495–510 (2006)
21. Wainer, A.L., Ingersoll, B.R.: The use of innovative computer technology for teaching social communication to individuals with autism spectrum disorders. *Res. Autism Spectrum Disord.* **5**, 96–107 (2011)
22. Coleman-Martin, M.B., Heller, K.W., Cihak, D.F., Irvine, K.L.: Using computer-assisted instruction and the nonverbal reading approach to teach word identification. *Focus Autism Other Dev. Disabil.* **20**, 80–90 (2005)

23. Hetzroni, O.E., Shalem, U.: From logos to orthographic symbols: a multilevel fading computer program for teaching nonverbal children with autism. *Focus Autism Other Dev. Disabil.* **20**, 201–212 (2005)
24. Knight, V.F., Wood, C.L., Spooner, F., Browder, D.M., O'Brien, C.P.: An exploratory study using science texts with students with autism spectrum disorder. *Focus Autism Other Dev. Disabil.* **30**, 86–99 (2015)
25. Escobedo, L., Nguyen, D., Boyd, L., Hirano, S., Rangel, A., Garcia-Rosas, D. et al.: MOSOCO: a mobile assistive tool to support children with autism practicing social skills in real-life situations, pp. 2589–2598 (2012)
26. Bauminger-Zviely, N., Eden, S., Zancanaro, M., Weiss, P.L., Gal, E.: Increasing social engagement in children with high-functioning autism spectrum disorder using collaborative technologies in the school environment. *Autism* **17**, 317–339 (2013)
27. Bernardini, S., Porayska-Pomsta, K., Smith, T.J.: ECHOES: an intelligent serious game for fostering social communication in children with autism. *Inf. Sci.* **264**, 41–60 (2014)
28. Cheng, Y.F., Huang, C.L., Yang, C.S.: Using a 3D immersive virtual environment system to enhance social understanding and social skills for children with autism spectrum disorders. *Focus Autism Other Dev. Disabil.* **30**, 222–236 (2015)
29. Chien, M.E., Jheng, C.M., Lin, N.M., Tang, H.H., Tael, P., Tseng, W.S., et al.: iCAN: a tablet-based pedagogical system for improving communication skills of children with autism. *Int. J. Hum Comput Stud.* **73**, 79–90 (2015)
30. Grynspan, O., Martin, J.-C., Nadel, J.: Multimedia interfaces for users with high functioning autism: an empirical investigation. *Int. J. Hum Comput Stud.* **66**, 628–639 (2008)
31. Josman, N., Ben-Chaim, H.M., Friedrich, S., Weiss, P.L.: Effectiveness of virtual reality for teaching street-crossing skills to children and adolescents with autism. *Int. J. Disabil. Hum. Dev.* **7**, 49–56 (2008)
32. Lahiri, U., Bekele, E., Dohrmann, E., Warren, Z., Sarkar, N.: Design of a virtual reality based adaptive response technology for children with autism. *IEEE Trans. Neural Syst. Rehabil. Eng.* **21**, 55–64 (2013)
33. Milne, M., Luerssen, M., Lewis, T., Leibbrandt, R., Powers, D.: Designing and evaluating interactive agents as social skills tutors for children with autism spectrum disorder (2011)
34. Rice, L.M., Wall, C.A., Fogel, A., Shic, F.: Computer-assisted face processing instruction improves emotion recognition, mentalizing, and social skills in students with ASD. *J. Autism Dev. Disord.* **45**, 2176–2186 (2015)
35. Swettenham, J.: Can children with autism be taught to understand false belief using computers? *J. Child Psychol. Psychiatry* **37**, 157–165 (1996)
36. Alves, S., Marques, A., Queirós, C., Orvalho, V.: Life is game prototype: a serious game about emotions for children with autism spectrum disorders. *PsychNology J.* **11**, 191–211 (2013)
37. Chen, C.H., Lee, I.J., Lin, L.Y.: Augmented reality-based self-facial modeling to promote the emotional expression and social skills of adolescents with autism spectrum disorders. *Res. Dev. Disabil.* **36**, 396–403 (2015)
38. LaCava, P.G., Golan, O., Baron-Cohen, S., Myles, B.S.: Using assistive technology to teach emotion recognition to students with Asperger syndrome a pilot study. *Remedial Spec. Educ.* **28**, 174–181 (2007)
39. Moore, D., Cheng, Y., McGrath, P., Powell, N.J.: Collaborative virtual environment technology for people with autism. *Focus Autism Dev. Disabil.* **20**, 231–243 (2005)
40. Tanaka, J.W., Wolf, J.M., Klaiman, C., Koenig, K., Cockburn, J., Herlihy, L., et al.: Using computerized games to teach face recognition skills to children with autism spectrum disorder: the let's face it! program. *J. Child Psychol. Psychiatry* **51**, 944–952 (2010)

Survey of String Similarity Approaches and the Challenging Faced by the Arabic Language

Sumayh S. Aljameel, James D. O'Shea, Keeley A. Crockett, Annabel Latham

Department of Computing, Math and Digital Technology

Manchester Metropolitan University

Sumayh.s.aljameel@stu.mmu.ac.uk, {j.d.oshea, k.crockett, A.latham}@mmu.ac.uk

Abstract—Measuring the similarity between strings plays an increasingly important role in many applications such as information retrieval, short answer grading, and conversational agent software. There has been much recent research interest in applying string similarity within Arabic language applications; however, the use of string similarity in Arabic poses a substantial challenge such as the complexity of the morphological system, ambiguity, and lack of resources. This survey discusses the existing research into string similarity approaches and the difficulties posed by the Arabic language by dividing them into three approaches; lexical-based similarity, semantic-based similarity, and hybrid similarity. The aim of this paper is to review these approaches and to identify suitable approaches with Arabic language.

Keywords—String Similarity; The Arabic Language; Lexical-Based Similarity; Semantic-Based Similarity; Hybrid Similarity; Arabic WordNet; Arabic Corpus.

I. INTRODUCTION

String similarity approaches play an important role in many applications within natural language processing (NLP) and related areas such as information retrieval [1], document clustering[2], essay scoring [3]and machine translation[4]. One of the more well-known areas where string similarity plays an important role is conversational agents (CA), which is a computer program that interacts with humans through natural language dialogue. O'Shea et al.[5]used string similarity in a conversational agent.

The similarity of words in the context of given string, can be evaluated in two ways: semantically and lexically. If two words have a similar sequence of characters, they are considered to be lexically similar. On the other hand, if the words have the same context and meaning although they contain different characters, they are considered to be semantically similar [6].

In this paper, lexical similarity is reviewed using character-based and statement-based algorithms [7], and semantic similarity is reviewed through Corpus-based and knowledge-based algorithms [8]. A furthermore, more recent-approach is reviewed in this paper; this is the hybrid similarity, which is a combination of different similarity measures. Figure 1 summarises these string similarity approaches.

The Arabic language is spoken and written by more than 300 million people in more than twenty countries worldwide [9].

Applying NLP tasks in general and in CA, in particular, is very challenging within the Arabic language. Arabic language has many features, which are considered as challenges for Arabic CA.

The first challenge is that there are 3 types of Arabic language, which are known as Classical, Modern and Colloquial. Classical Arabic, which is used in the Quran, is more complex in its grammar and vocabulary than modern Arabic. It has a large number of diacritics that facilitate word pronunciation and detection in their grammatical cases. The second type is the Modern Arabic, all diacritics were omitted to make the process of the reading and writing easier and fast. This type is considered as the official language of the Arab countries and is used in everyday language, education, and in the media. Usually, computational Arabic-based researches use Modern Arabic[10]. In Colloquial Arabic, which is the third type, the grammar and vocabulary are less sophisticated in comparison to Modern Arabic. However, most people use it in their everyday spoken conversations and in informally written letters because of its simplicity [11]. Arabic people make many mistakes in grammar when they use Modern Arabic and they mix between Modern and Colloquial Arabic. Moreover, different Arab countries have different Colloquial Arab. Consequently, the challenges increase for Arabic CA to recognise the user's utterance. The second challenge is the Arabic Morphology. Arabic language is complex because of the morphological variation and the agglutination phenomenon[12]. The form of letters change depending on their position in the word. In addition, the word may consist of prefixes, lemma and suffixes in different combinations, which results in a very complicated morphology[13]. These features will increase the challenge for Arabic CA to determine the grammatical case for a user's utterance words. The third challenge is the capitalisation. Arabic language does not support capitalisation of proper nouns such as countries names, people names. Whereas, in Latin languages these begin with a capital letter[14]. Arabic CA may not recognise these words and this increases the difficulty of detecting such nouns in textual Arabic conversations.

The fourth challenge is that the short vowels and diacritics are optional in Arabic. They are needed for pronunciation and disambiguation. However, Modern Arabic texts do not include

diacritics. Therefore, the form of a word may have two or more different meanings depending on the context. This creating an ambiguity during the conversation in Arabic CA.

The last challenge is the lack of linguistic resources. Generally, there is a limitation on the number of Arabic linguistic resources, which are available free for research purposes. This challenge is discussed in section III.

The previous challenges need to be resolved when building a system for Arabic CA. The aim of this paper is to review some approaches of String similarity (Lexical and Semantic) to use it with Arabic applications such as CA. The remainder of this paper is set out as follows: Section II presents the lexical-based similarity measurements (LBS) offered by the two approaches of character-based similarity and statement-based similarity; Section III discusses semantic-based similarity measurement (SBS) from the two aspects of corpus-based similarity and knowledge-based similarity; Section IV describes the hybrid similarity; and Section V presents the conclusion.

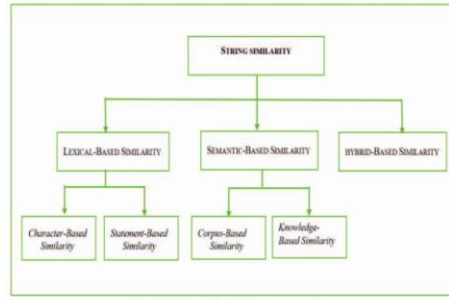


Fig 1. String Similarity Approaches

II. LEXICAL-BASED SIMILARITY

Lexical-based similarity approaches (LBS) [15] provide the similarity on the basis of character and statement matching. These methods are based on measuring the distance between two strings to identify the similarity between them [6]. In this paper, LBS measures are divided into two categories: character-based and statement-based distance measures. Two different algorithms are described under each of these two categories.

A. Character-Based Similarity

In this section, the similarity measurements are introduced, based on the distance between each character in two strings. Character-based similarity measurements are designed to handle typographical errors[16].

1. Levenshtein distance similarity

The Levenshtein distance [17] uses a distance factor to calculate the similarity between two given strings. The distance is computed by counting the minimum number of operations needed to transform one string into the other, with possible operations being insertion, substitution or deletion of a single

character. The Levenshtein distance between two strings a and b is given by

$$\text{lev}_{a,b}(a, b) = \begin{cases} \max(i, j) & \text{if } \min(i, j) = 0 \\ \min \begin{cases} \text{lev}_{a,b}(i-1, j) + 1 \\ \text{lev}_{a,b}(i, j-1) + 1 \\ \text{lev}_{a,b}(i-1, j-1) + 1 \end{cases} & \text{otherwise} \end{cases} \quad (1)$$

where i, j are the indices for strings a and b respectively [18]. It has been reported [19] that the Levenshtein distance similarity generates improved results when it is used for a short string; the long string cost of the Levenshtein distance algorithm is same as the length of the string. For Arabic language, many researchers used Levenshtein distance algorithm such as Shaalan et al. [20] who used it to develop Spell Checking tool for Arabic words. However, Levenshtein does not give accurate results when applied on the Arabic language. For example, when comparing (الحمد) and (الحمد), the Levenshtein distance is equal to one. The Arabic reader does not differentiate between (Alif with Hamza) and (Alif without Hamza) and therefore the distance should be zero [21].

2. N-gram similarity

The n-gram method [22] determines the similarity of a subsequence of n items from a given text sequence. The similarity is computed based on a distance between each character in two strings. The distance is calculated by dividing the number of similar grams by a maximal number of n-grams [18]. For different words w, w_1, w_2, w_3, \dots , N-gram similarity is:

$$P(w) = P(w_1)P(w_2/w_1)P(w_3/w_2, w_1) \dots \dots \dots \quad (2)$$

N-gram is used in many areas of computer science, for instance in the design of kernels, which are developed using a machine learning algorithm such as the support vector machine to generate learning from string data. It has been shown [23] that the n-gram similarity measure is a high performance algorithm, although its accuracy is somewhat low. The work in [24] used an N-gram conflation method to convert a word into a sequence of N-grams and apply it within the context of Arabic textual retrieval systems. The study indicates that the N-gram approach does not appear to provide an efficient conflation approach within Arabic context.

B. Statement-Based Similarity

In this section, the similarity measurements are introduced, based on the distance between each word in two strings. Statement-Based Similarity measurements are also effective for solving the typographical errors. However, Character-based similarity measures fail to capture the similarity with word order problem (e.g., "John Smith" versus "Smith, John"). Statement-based similarity measures try to compensate for this problem[16].

1. Cosine similarity

Cosine similarity [25] determines the similarity between two pieces of text by representing each piece of text in the form of vector. A dimension in Euclidean space is defined by each word in the text. The value in that dimension corresponds to the frequency of each word.

For two texts t_1 and t_2 the cosine similarity is:

$$SIM(t_1, t_2) = \frac{\sum_{i=1}^n t_{1i} t_{2i}}{\sqrt{\sum_{i=1}^n t_{1i}^2} \times \sqrt{\sum_{i=1}^n t_{2i}^2}} \quad (3)$$

For Arabic language, Al-Ramahi et al. [26] used cosine similarity to develop a system that determines the similarity between course descriptions of the same subject for credit transfer among various universities or similar academic programs. The results indicate that this technique has demonstrated better performance comparing with others techniques such as Dice's Similarity.

Another work in [27] studied the different lexical similarity measures in Arabic information retrieval and it is found that the Cosine Similarity Measure is the best measure comparing with others measures, which are: Inner Product, Dice Coefficient, Jaccard Coefficient, Inclusion Similarity Coefficient, Overlap Coefficient Measure, Euclidean distance Measure and Manhattan Distance Measure.

2. Web Jaccard Similarity

The Web Jaccard similarity [28] identifies the similarity between words by calculating the number of elements within the intersection set and dividing this by the number of elements in the union set.

$$Web\ Jaccard(P, Q) = \begin{cases} 0 & \text{if } H(P \cap Q) \leq C \\ \frac{H(P \cap Q)}{H(P) + H(Q) - H(P \cap Q)} & \text{otherwise} \end{cases} \quad (4)$$

where C is a predefined threshold, and $H(P)$ and $H(Q)$ are the page counts for queries P and Q respectively.

Khafajeh et al. [29] designed an automatic Arabic thesauri using term-term similarity. They compared Jaccard similarity measure with other measures such as Cosine, Dice and Inner product. The results indicate the Jaccard and Dice similarity measurements have the same performance, while the Cosine and Inner similarity measurements are also similar; however, they are slightly more efficient than Jaccard and Dice measures.

III. SEMANTIC-BASED SIMILARITY

Semantic similarity determines the similarity between strings based on their meaning rather than using character-matching [30]. Semantic similarity uses two main approaches to compute the similarity between strings: corpus-based and knowledge-based measures.

A. Corpus-Based Similarity

Corpus-based similarity [31] determines the similarity between words by using information from a large corpus, that is, a large body of different forms of text for comparison purposes.

Arabic is a poorly resourced language for corpus-based approaches in comparison with English, since there is a lack of sufficient data, and this negatively affects research into corpus-based and natural language processing in Arabic. More recently, a number of Arabic corpora have been developed; however, little overall improvement in the overall situation has been made [32].

Al-Thubaity [33] reviewed fourteen Arabic corpora and categorised them by their target language, designated purpose, text date, location, text domain, representativeness, mode of text, size, text type/medium, and balance. The paper reported some limitations of this review. Several of these corpora do not provide any information regarding the time period covered by the texts. In addition, the size of the reviewed corpora is large, such as the arabiCorpus, the arTenTen12 corpus, and the Leeds Internet Arabic Corpora. Moreover, for all corpora the component texts are not classified with respect to their dates or the time period to which they belong; there is therefore a limitation on the usability of the corpus and a difficulty in comparing languages used in different periods, and in observing how the Arabic language has evolved. For example, texts collected more recently are more likely to be written in Modern Arabic language, while texts collected from the old period are more likely to be written in classical Arabic language. Consequently, the efficiency of corpus-based similarity measurements will be affected because of the different texts language types. The design criteria for many of the corpora are unclear, which makes it difficult to evaluate accurately the findings of any research based on them, and there is no justification for excluding such criteria.

A well-known early measurement technique, which utilise a corpus is, latent semantic similarity (LSA) [34]. It is defined as the similarity in meaning between two words. A matrix is formed in the LSA measure in which the row represents a particular word and the column represents the document or paragraph. To reduce the size of this matrix, the singular value matrix (SVM) approach can be used. After forming the matrix, the cosine similarity or Web Jaccard similarity algorithms can be applied to obtain the similarity between the words. LSA similarity is relatively fast in comparison to other dimensionality reduction models[35]. The main disadvantage of this measure is that the similarity does not use any syntactic information from the compared texts. For example, the sentences "The dog hunted the man" and "The man hunted the dog" will be considered as identical [36]. Refaat et al. [37] developed an automatic Arabic essay scoring system using LSA similarity measurement. They improved LSA by processing the entered text before applying the measure. The new algorithm unifies the form of letters, deletes the formatting, replaced synonyms, and deletes "Stop Words" in order to create a matrix that represents texts more efficient than the general form of LSA matrix. The result indicates that the correlation between the human assessor and the system is 0.91, which is good in

comparison with the human assessor and the traditional LSA that has a correlation varied from 0.78 to 0.87.

B. Knowledge-Based Similarity

Knowledge-based similarity [38] is a semantic-based similarity which determines the similarity between words using information derived from semantic networks. The most commonly-used semantic networks are WordNet [39] and the Natural Language Toolkit (NLTK) [40]. WordNet is a database that collects words into sets of synonyms called synsets; for each synset, WordNet provides short definitions and usage examples, and records a number of relations, which include hyponyms, hypernyms, coordinate terms, holonyms and meronyms [41].

WordNet is used in various areas such as information retrieval [42], semantic similarity [43], and word sense disambiguation [44]. Due to the success of WordNet in English language applications, several projects are currently being conducted to develop WordNet for other languages [45]. The Global WordNet Organisation, a non-profit public group, was established to support the development of WordNet and the production of dictionaries for other languages based on the English version of WordNet. It provides connectivity between the WordNet dictionaries in various languages such as Arabic, Persian, Hebrew, Albanian and the African and Indian languages [46].

Arabic WordNet (AWN) has been developed using the same methodology as EuroWordNet. It consists of 11,270 synsets and contains 23,496 Arabic expressions (words and multiwords) [47]. Numerous researchers have used AWN in their work, for example within an Arabic Information Retrieval system [48] and an Arabic Question Answering (Q/A) system [49]. The major limitations of the current AWN are a lack of information and concepts in comparison with WordNet in the English language, and some semantic relations between synsets. Many Arabic concepts have not been included in the AWN database. This limitation forms a major impediment for using AWN as the source for knowledge-based approaches. AWN could be enhanced and extended through several different approaches, for example the addition of new synsets, improvements to the existing synsets, improvements to the hyponym/hypernym relations, and the use of named entity (NE) synsets, 'glosses' and verb categorisations [50].

Several researchers have attempted to extend AWN through the addition of named entity synsets in order to enhance the performance of AWN. Al Khalifa and Rodriguez [47] used Arabic Wikipedia to extract Arabic Named Entities (NEs) automatically and attach them to the Arabic WordNet. Abouenour et al. [50] focused on automated AWN improvement for the NE synsets, by using AWN in Q/A systems to improve system performance. However, these researchers reported that the coverage, hierarchy and relations associated with the NE synsets within AWN require improvement. They therefore proposed an approach based on the Yago ontology, a freely available ontology with a large coverage of NE. Their approach is based on the addition of both new NE and their super-types. In addition, Saif et al. [51] have

mapped an Arabic WordNet synset to its corresponding article in Wikipedia, using monolingual and bilingual features to overcome the lack of semantic information in Arabic WordNet. This mapped method increased the coverage of Arabic WordNet by inserting new synsets from Wikipedia. Apart from this work focusing on NE synsets, the coverage of semantic relations is rather limited and needs to be developed.

A few researchers have attempted to overcome this limitation, such as Boudabous et al. [52] who suggested improving the performance of AWN using a linguistic method based on morpho-lexical patterns in order to add semantic relations between synsets using Arabic Wikipedia. However, the size of the Arabic Wikipedia remains an order of magnitude smaller than the English version, while the structured data in the Arabic version of Wikipedia, such as information boxes, are of lower quality on average in terms of coverage and accuracy. Consequently, the semantic similarity area using AWN needs further research in order to be more reliable. It is not sufficient enough to rely on AWN to apply semantic similarity measurements and to find the similarity between texts.

IV. HYBRID-BASED SIMILARITY

Hybrid methods use multiple similarity measures, and have been the subject of much research. Mihalcea et al. [8] tested eight semantic similarity measures, of which six were knowledge-based and two were corpus-based measures. These researchers evaluated the eight algorithms separately, and then combined them together, and found that the most efficient performance was achieved when several similarity metrics were combined into one. Kaleem et al. [53] used a hybrid approach which combined a lexical sentence similarity measure with pattern-matching techniques in order to calculate the matching strength between the user's expressions and the scripted patterns in Urdu language. The similarity algorithm comprises of Levenshtein algorithm to calculate the similarity between two strings, while Bipartite Matching [54] is used to determine the word order variation and Kuhn-Munkres algorithm [55], is used to find the maximum sum of a given matrix of weights. This hybrid algorithm overcomes the challenges of string similarity and free word order in Urdu language or other languages with free word order such as Arabic language. Gomaa et al. [7] used lexical-based and corpus-based similarity measures to develop their Short Answer Grading system. They tested the measurements using three stages. In the first stage, they assessed the similarity between the model answer and student answer using thirteen lexical-based measurements, of which seven were statement-based and six were character-based measures. Their results showed that the best correlation values achieved using character-based and statement-based measures were obtained using the n-gram and block distance (Manhattan distance) approaches respectively. In the second stage, they measured the similarity using corpus-based similarity measurements [56]: DISCO1 (Computes the similarity of the first order between two words based on their collocation sets) and DISCO2 (Computes similarity of the second order between two words based on their sets of distributing similar words). The results

showed that DISCO1 achieved more efficient correlation values. In the third stage, the similarity was evaluated by combining lexical-based and corpus-based measures. The best correlation value was obtained from mixing n-gram with DISCO1 similarity techniques.

V. CONCLUSION

This paper presents a survey of three types of similarity measurements and gives an overview of the challenges of using similarity approaches in the Arabic language. The three types of string similarity approaches covered in this survey are lexical-based, semantic-based and hybrid-based methods. Lexical-based similarity approaches measure the similarity between texts based on character matching. Four algorithms of this form were summarised: two character-based and two statement-based algorithms. The second type of approach is semantic-based similarity, which determines the similarity between pieces of text based on their meaning. The various Arabic corpora used in this type of approach, and Arabic WordNet (AWN), were reviewed and discussed. The third approach is hybrid-based similarity, which combines together multiple similarity measurements.

The use of similarity approaches within the Arabic language poses significant challenges.

1. Arabic is a complex language, which is often ambiguous in nature.
2. Currently, AWN lacks the necessary information and a range of concepts in comparison with the English language WordNet, and some semantic relations between synsets are missing. Moreover,
3. Moreover, most of the Arabic corpora does not cover all possible domains and words, with each corpus being focused on a specific domain and lacking information.

In concluding, the Cosine similarity measurement was used in many Arabic systems and the results show a more efficient performance compared to other lexical measurements. However, using the lexical similarity approach is not reliable due to the different features in Arabic language such as morphology. The semantic similarity approach is not reliable with Arabic language because of the weakness of AWN and Arabic Corpora. In most recent times, the hybrid similarity approach is considered as a promising approach with Arabic language because it combines more than one type of measurements which leads to the similarity being more robust. Further research needs to be carried out in order to enhance both AWN and Arabic corpora in order to use it with semantic measurements. Moreover, Future works will be carried out to conduct an experiment that evaluates the similarity measures mentioned in this paper with Arabic languages texts.

REFERENCES

- [1] J. Singh, P. Singh, and Y. Chaba, "A Study of Similarity Functions Used in Textual Information Retrieval in Wide Area Networks," *International journal of computer science and information technologies*, vol. 5, no. 6, pp. 7880-7884, 2014.
- [2] B. Ramesh and M. W. Watzke, "Clustering strings using N-grams," ed: Google Patents, 2010.
- [3] R. Mezher and N. Omar, "A Hybrid Method of Syntactic Feature and Latent Semantic Analysis for Automatic Arabic Essay Scoring," *Journal of Applied Sciences*, vol. 16, no. 5, p. 209, 2016.
- [4] M. Hanumanthappa, S. Rashmi, and M. V. Reddy, "Metrics for evaluating phonetics machine translation in Natural Language Processing through modified Edit Distance algorithm-A naïve approach," in *Computer Communication and Informatics (ICCCI), 2015 International Conference on*, 2015, pp. 1-7: IEEE.
- [5] K. O'Shea, Z. Bandar, and K. Crockett, "A conversational agent framework using semantic analysis," *International Journal of Intelligent Computing Research (IJICR)*, vol. 1, no. 1/2, 2010.
- [6] W. H. Gomaa and A. A. Fahmy, "A survey of text similarity approaches," *International Journal of Computer Applications*, vol. 68, no. 13, 2013.
- [7] W. H. Gomaa and A. A. Fahmy, "Short answer grading using string similarity and corpus-based similarity," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 3, no. 11, 2012.
- [8] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," in *AAAI*, 2006, vol. 6, pp. 775-780.
- [9] K. Versteegh, *The arabic language*. Edinburgh University Press, 2014.
- [10] N. Y. Habash, "Introduction to Arabic natural language processing," *Synthesis Lectures on Human Language Technologies*, vol. 3, no. 1, pp. 1-187, 2010.
- [11] E. Al-Saidat and I. Al-Momani, "Future markers in Modern Standard Arabic and Jordanian Arabic: a contrastive study," *European Journal of Social Sciences*, vol. 12, no. 3, pp. 397-408, 2010.
- [12] Y. Kadri and A. Benyamina, "A syntax semantic analyzer for Arabic language," University of Oran, 1992.
- [13] S. Abuleil, K. Alsamara, and M. Evens, "Acquisition system for Arabic noun morphology," in *Proceedings of the ACL-02 workshop on Computational approaches to semitic languages*, 2002, pp. 1-8: Association for Computational Linguistics.
- [14] S. Alanazi, B. Sharp, and C. Stanier, "A Named Entity Recognition System Applied to Arabic Text in the Medical Domain," *International Journal of Computer Science Issues (IJCSI)*, vol. 12, no. 3, p. 109, 2015.
- [15] C. R. Rensch, "Calculating lexical similarity," *Windows on bilingualism*, pp. 13-15, 1992.
- [16] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate record detection: A survey,"

- IEEE Transactions on knowledge and data engineering*, vol. 19, no. 1, pp. 1-16, 2007.
- [17] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," in *Soviet physics doklady*, 1966, vol. 10, p. 707.
- [18] N. Pradhan, M. Gyanchandani, and R. Wadhvani, "A Review on Text Similarity Technique used in IR and its Application," *International Journal of Computer Applications*, vol. 120, no. 9, 2015.
- [19] K. U. Schulz and S. Mihov, "Fast string correction with Levenshtein automata," *International Journal on Document Analysis and Recognition*, vol. 5, no. 1, pp. 67-85, 2002.
- [20] K. F. Shaalan, M. Attia, P. Pecina, Y. Samih, and J. van Genabith, "Arabic Word Generation and Modelling for Spell Checking," in *LREC*, 2012, pp. 719-725.
- [21] H. H. Abdel Ghafour, A. El-Bastawissy, and A. F. A. Heggazy, "AEDA: Arabic edit distance algorithm Towards a new approach for Arabic name matching," in *Computer Engineering & Systems (ICCES), 2011 International Conference on*, 2011, pp. 307-311: IEEE.
- [22] M. Francis and R. N. Kn, "An Algorithm for Plagiarism Detection in Malayalam Language Documents Using Modified n-gram Model," 2014.
- [23] A. Barrón-Cedeno, P. Rosso, E. Agirre, and G. Labaka, "Plagiarism detection across distant language pairs," in *Proceedings of the 23rd International Conference on Computational Linguistics*, 2010, pp. 37-45: Association for Computational Linguistics.
- [24] S. H. Mustafa and Q. A. Al-Radaideh, "Using N-grams for Arabic text searching," *Journal of the American Society for Information Science and Technology*, vol. 55, no. 11, pp. 1002-1007, 2004.
- [25] G. Qian, S. Sural, Y. Gu, and S. Pramanik, "Similarity between Euclidean and cosine angle distance for nearest neighbor queries," in *Proceedings of the 2004 ACM symposium on Applied computing*, 2004, pp. 1232-1237: ACM.
- [26] M. A. Al-Ramahi and S. H. Mustafa, "N-Gram-Based Techniques for Arabic Text Document Matching: Case Study: Courses Accreditation," 2012.
- [27] I. Hajeer, "Comparison on the Effectiveness of Different Statistical Similarity Measures," *International Journal of Computer Applications*, vol. 53, no. 8, 2012.
- [28] P. Jaccard, *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*. Impr. Corbaz, 1901.
- [29] H. Khafajeh, N. Yousef, and G. Kanaan, "Automatic query expansion for arabic text retrieval based on association and similarity thesaurus," in *Proceedings the European, Mediterranean & Middle Eastern Conference on Information Systems (EMCIS), Abu Dhabi, UAE*, 2010.
- [30] Y. Li, Z. A. Bandar, and D. McLean, "An approach for measuring semantic similarity between words using multiple information sources," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 15, no. 4, pp. 871-882, 2003.
- [31] A. Islam and D. Inkpen, "Semantic text similarity using corpus-based word similarity and string similarity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 2, no. 2, p. 10, 2008.
- [32] M. M. A. Alqahtani and E. Atwell, "A Review of Semantic Search Methods to Retrieve Information from the Qur'an Corpus," 2015.
- [33] A. O. Al-Thubaity, "A 700M+ Arabic corpus: KACST Arabic corpus design and construction," *Language Resources and Evaluation*, vol. 49, no. 3, pp. 721-751, 2015.
- [34] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse processes*, vol. 25, no. 2-3, pp. 259-284, 1998.
- [35] D. Tomás, J. L. V. González, E. Bisbal, and L. Moreno, "Experiments with LSA for Passage Re-Ranking in Question Answering," in *CLEF (Working Notes)*, 2006: Citeseer.
- [36] F. Al-marsoomi, "The Development of a Framework for Semantic Similarity Measures for the Arabic Language," PhD, Computer Science, Manchester Metropolitan University, UK, 2015.
- [37] M. M. Reafat, A. A. Ewees, M. M. Eisa, and A. Ab Sallam, "Automated assessment of students arabic free-text answers," *Int J Cooperative Inform Syst*, vol. 12, pp. 213-222, 2012.
- [38] A. Budanitsky and G. Hirst, "Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures," in *Workshop on WordNet and Other Lexical Resources*, 2001, vol. 2, p. 2.2.
- [39] G. A. Miller, "WordNet: a lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39-41, 1995.
- [40] E. Loper and S. Bird, "NLTK: The natural language toolkit," in *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*, 2002, pp. 63-70: Association for Computational Linguistics.
- [41] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to wordnet: An on-line lexical database*," *International journal of lexicography*, vol. 3, no. 4, pp. 235-244, 1990.
- [42] S. Kara, Ö. Alan, O. Sabuncu, S. Akpınar, N. K. Cicekli, and F. N. Alpaslan, "An ontology-based retrieval system using semantic indexing," *Information Systems*, vol. 37, no. 4, pp. 294-305, 2012.
- [43] M. Batet, D. Sánchez, and A. Valls, "An ontology-based measure to compute semantic similarity in

- biomedicine," *Journal of biomedical informatics*, vol. 44, no. 1, pp. 118-125, 2011.
- [44] Y. Gutiérrez, A. Fernández, A. Montoyo, and S. Vázquez, "UMCC-DLSI: Integrative resource for disambiguation task," in *Proceedings of the 5th International Workshop on Semantic Evaluation*, 2010, pp. 427-432: Association for Computational Linguistics.
- [45] A. M. Al-Zoghby, A. S. E. Ahmed, and T. T. Hamza, "Arabic Semantic Web Applications—A Survey," *Journal of Emerging Technologies in Web Intelligence*, vol. 5, no. 1, pp. 52-69, 2013.
- [46] (2016). <http://globalwordnet.org/wordnets-in-the-world>. Available: <http://globalwordnet.org/wordnets-in-the-world>
- [47] M. Alkhalifa and H. Rodríguez, "Automatically extending named entities coverage of Arabic WordNet using Wikipedia," *International Journal on Information and Communication Technologies*, vol. 3, no. 3, 2010.
- [48] M. A. Abderrahim, M. E. A. Abderrahim, and M. A. Chikh, "Using Arabic wordnet for semantic indexation in information retrieval system," *arXiv preprint arXiv:1306.2499*, 2013.
- [49] I. AlAgha and A. Abu-Taha, "AR2SPARQL: An Arabic Natural Language Interface for the Semantic Web," *International Journal of Computer Applications*, vol. 125, no. 6, 2015.
- [50] L. Abouenour, K. Bouzoubaa, and P. Rosso, "Using the Yago ontology as a resource for the enrichment of Named Entities in Arabic WordNet," in *Proceedings of The Seventh International Conference on Language Resources and Evaluation (LREC 2010) Workshop on Language Resources and Human Language Technology for Semitic Languages*, 2010, pp. 27-31.
- [51] A. Saif, M. J. Ab Aziz, and N. Omar, "Mapping Arabic WordNet synsets to Wikipedia articles using monolingual and bilingual features," *Natural Language Engineering*, pp. 1-39, 2015.
- [52] M. M. Boudabous, N. C. Kammoun, N. Khedher, L. H. Belguith, and F. Sadat, "Arabic WordNet semantic relations enrichment through morpho-lexical patterns," in *Communications, Signal Processing, and their Applications (ICCSPA), 2013 1st International Conference on*, 2013, pp. 1-6: IEEE.
- [53] M. Kaleem, J. D. O'Shea, and K. A. Crockett, "Word order variation and string similarity algorithm to reduce pattern scripting in pattern matching conversational agents," 2014, pp. 1-8: IEEE.
- [54] D. Dasgupta *et al.*, "A comparison of multiobjective evolutionary algorithms with informed initialization and kuhn-munkres algorithm for the sailor assignment problem," in *Proceedings of the 10th annual conference companion on Genetic and evolutionary computation*, 2008, pp. 2129-2134: ACM.
- [55] R. E. Burkard and E. Cela, "Linear assignment problems and extensions," in *Handbook of combinatorial optimization*: Springer, 1999, pp. 75-149.
- [56] P. Kolb, "Disco: A multilingual database of distributionally similar words," *Proceedings of KONVENS-2008, Berlin*, 2008.

Development of An Arabic Conversational Intelligent Tutoring System for Education of Children with ASD

Sumayh S. Aljameel, James D. O'Shea, Keeley A. Crockett, Annabel Latham, Mohammad Kaleem

Department of Computing, Math and Digital Technology

Manchester Metropolitan University

Sumayh.s.aljameel@stu.mmu.ac.uk, {j.d.oshea, k.crockett, a.latham, m.kaleem}@mmu.ac.uk

Abstract—This paper presents a novel Arabic Conversational Intelligent Tutoring System (CITS) that adapts the learning styles VAK for autistic children to enhance their learning. The proposed CITS architecture uses a combination of Arabic Pattern Matching and Arabic Short Text Similarity to extract the responses from the resources. The new Arabic CITS, known as LANA, is aimed at children with autism (10 to 16 years old) who have reached a basic competency with the mechanics of Arabic writing. This paper describes the architecture of LANA and its components. The experimental methodology is explained in order to conduct a pilot study in future.

Keywords— Autism Spectrum Disorder; Conversational Agent; Intelligent Tutoring System; String Similarity; The Arabic Language.

I. INTRODUCTION

Autism spectrum disorder (ASD) has four main diagnostic categories: Autistic Disorder, Asperger Syndrome, Pervasive Developmental Disorder and Childhood Disintegrative Disorder. This project has been specifically designed for high-functioning individuals or Asperger's (i.e. those with higher verbal IQ). People with ASD have the ability to deal with technology and computers better than neurotypical people because of the structured environment that is offered by software. This environment is preferred by people with autism because it is more in-line with their routine and repetitive behaviour [1][2][3]. Traditional education using human tutors is a challenge for children with autism. Whereas virtual tutors can meet the individual children's needs. One advantage of using a virtual tutor is that some autistic children are uncomfortable looking at faces and human interaction. A virtual tutor's face can be made to look like a cartoon, which is less intrusive and more familiar to the child, and when the child becomes a custom with the virtual tutor, the complexity can be increased. There is some evidence to suggest that autistic children who were taught by virtual tutors gained more information compared to those who were taught by human tutors [4]. Another advantage is, the virtual tutor allows children to learn and practice their skills independently. Consequently, the human tutor can focus on the complex aspects and tasks while the software covers the general topics [5][6]. Children with autism tend to learn through one of three kinds of learning styles (VAK): through seeing (Visually),

through hearing (Auditory), and through touching an object (Kinesthetically). Children with autism are more likely to rely on one style of learning, which is determined either by observation by a teacher or by using one of a number of behavioral questionnaires [7]. The most common of which is "VAK Learning Styles Self-Assessment Questionnaire" [8].

According to Papanikolaou et al.[32], Intelligent Tutoring Systems (ITS) are computerised learning systems that mimic human tutors to provide more personalised learning than previous content delivery systems. A Conversational Intelligent Tutoring Systems (CITS) is an extension of ITS. It integrates natural language interfaces rather than menus to allow users to learn topics through discussion as they would in the classroom. Many researchers developed CITS for different domains. AutoTutor [33] is one example of CITS; it is a computer tutor that is designed to assist students at university level to learn an introductory computer literacy course. The system teaches the fundamentals of hardware, operating systems, and the Internet. However, AutoTutor does not consider student learning styles during a tutoring session. This means that all students are provided with the same learning experience regardless their abilities. Oscar [34] is another example of CITS, which overcomes AutoTutor's limitations. Oscar leads a tutoring conversation and it dynamically predicts and adapts to a student's learning style. Oscar was developed for students at University level and teaches SQL. In the Arabic language, CITS is considered as a new area of research. Recently, Alobaidi et al. [15] developed an Arabic CITS (Abdullah) for modern Islamic education. Abdullah CITS is an online system that aims to teach children aged 10 to 12 years old the essential topics in Islam. The system allows conversation, discussion and interpretation with verses in classical Arabic language by engaging in dialogue using Modern Arabic language. To our knowledge, no academic research exists on Arabic conversational intelligent tutoring systems for Autistic children.

In this research, an Arabic CITS was developed called LANA. The intention behind the research and development of LANA was to support education for autistic children (10 to 16 years old) who have reached a basic competency with the mechanics of Arabic writing. The LANA CITS engages children with a science tutorial delivered in modern standard Arabic. The

curriculum material that is delivered to students through LANA is mapped to the VAK model. The LANA architecture utilises pattern matching (PM) and Arabic short text similarity (STS) to extract the responses to user utterances from structured stored resources in particular domain (science). This project measures learning gain to evaluate the ability of a CITS to adapt to autistic children's learning styles. The current phase of the research focuses on two research questions:

1. Can an Arabic Conversational Intelligent Tutoring System (CITS) adapt to the learning styles of autistic children?
2. Can an Arabic Conversational Intelligent Tutoring System (CITS) for children with autism enhance their learning?

This paper is organised as follows: Section II describes the Conversational Agent and its approach. Section III contains the LANA architecture and its components. Section IV discusses the experimental methodology. Section V includes the conclusion and future works.

II. CONVERSATIONAL AGENT

A. Background

A Conversational Agent (CA) is defined as 'a computer program which interacts with a user through natural language dialogue and provides some form of service' [9]. According to Crockett et al. [10], A CA is defined as having 'the ability to reason and pursue a course of action based on its interactions with humans and other agents'. CAs have become a popular communication method and have been used in many applications in different domains such as e-commerce, medication, and education [11] [12]. Also, CAs are used effectively in many application such as student's debt management guidance [13], database interfacing [14], and as student's assistant by providing problem-solving advice as they learn [15]. There are two categories of CAs, 'Embodied CAs' (ECA) and linguistic CA's (LCA). Both categories are becoming increasingly popular in the field of CA. An embodied conversational agent (ECA) is defined as an agent which simulated face-to-face conversation with the user, by presenting the user with an avatar who talks to the user and uses the naturalistic modes of communication such as hand, head, facial expressions and body gestures [16]. A Linguistic conversational Agent (LCA) is defined as an agent who can converts with the user in written or spoken form [17].

B. Conversation Agent Approaches

There are three main approaches that are used to develop linguistic based CAs. These are Natural Language Processing (NLP); Pattern Matching (PM); and semantic sentence similarity measures (SSM), discussed briefly below.

Natural language processing (NLP) is defined as "the computational processing of textual materials in natural human languages" [18]. Researchers have been encourage to develop CAs based on NLP, focusing on extracting the utterance using grammar rules and list of attribute pairs from the conversation. This information is used to fill a template-based structure to generate a suitable response. Nevertheless, NLP based CAs emphasis on structured lexicon, parsing and various other language handling aspects without reasoning about the meaning

of the utterance. According to Demir [19], an NLP-based CA should understand the user's utterance in order to produce human-like conversations, meaning the semantic content of expressed information should be recognized. There are certain other disadvantages of NLP-based CAs. First, the system might not recognise the misspelling of a word in the conversation and so the system may fail to understand the whole conversation. Second, NLP-based CAs may not comprehend dialogue due to natural mistakes in grammar that people often make in everyday conversation. Thirdly, natural language is very ambiguous due to the rich in form and structure, which means that the word might have more than one meaning (lexical ambiguity) or a sentence might have more than one structure (syntactic ambiguity) [20].

Text-based Pattern Matching (PM) is the process of measuring the degree of matching between a string or sequence of strings and a piece of text [21]. From a CA perspective, PM algorithms are used to handle user conversations by matching a CA's patterns against a user's utterance. A typical pattern is generally composed of wildcards, words and spaces. A wildcard is a symbol used to match a portion of the user's utterance [22]. InfoChat [23] as well as ALICE [24] are two examples of CAs incorporating PM algorithms. The typical procedure for such CAs is organizing a scripted domain into contexts (about a particular topic), each of which has a number of related rules. Each rule has structured patterns that represent a user's utterance. PM is considered one of the more successful methods for developing CAs [25]. PM is deemed to have various benefits. For example, it has reduced operating and computing costs, as well as requiring less time for processing, due to there being no need for intricate procedures prior to processing. Furthermore, PM is effective regardless of language and is straightforward to comprehend. As a result, web-based interactions and conversations that are happening live and involve many individuals can be analysed and aided through CAs based on PM method [26]. An additional advantage is that PM CAs are able to overcome many linguistic challenges, which the NLP approach faces, such as morphological changes occurring on words in the form of prefixes and suffixes. This is largely addressed through the utilisation of pattern wildcards that can amend mistakes in spelling and grammar, as well morphology amendments [27].

The main limitation of using PM based CAs is the needs for a large number of patterns to implement a coherent domain. This limitation causes many issues. For example, certain keywords may have multiple meanings as well as singular and plural configurations, thus a programmer will have to take in to account these various possibilities and innate linguistic depth in terms of morphology. If one were dealing with Arabic language, then its various morphological amendments, prefixes and suffixes can be compensated for utilising various methods, however the PM procedure can be particularly complex, especially in scripting. A further issue is that a user utterance may be written in multiple ways through adopting various alternative words and phrase structure.

Information, philosophy and linguistic theories have all been adopted as a means of assessing sentence similarity, which is

the third approach for CAs [28]. Generally, similarity as a term is used to describe the similarity level between two objects. Semantics is “the study of the meaning of linguistic expressions” [29]. Information retrieval, question answering, information extraction machine translation, data mining machine translation, and conversational agents have all been aided by means of similarity-based research, while it also has wider application to AI and computer processing as a whole. Applying the sentence similarity technique in CA development is more effective than other techniques because it replaces the scripted patterns by a few natural language sentences in each rule. Consequently, the scripting effort and script size are reduced to a minimum. Thus, there is a reduction in the maintenance cost of such scripts [30] [31]. One limitation of using this approach is a lack of previous research in this area for Arabic language.

III. LANA CITS ARCHITECTURE

This section illustrates the design and development of LANA CITS architecture. LANA CITS architecture was designed based on three main components as shown in Figure 1, which are: component 1 (consisting of the system knowledge base which includes: the tutorial scripts, log file, user's profile, and general contexts), component 2 (The Arabic CA), and component 3 (the ITS). The following sections will describe each component in detail.

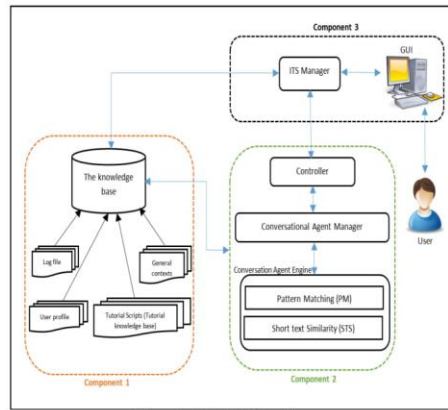


Figure 1: LANA CITS architecture

A. Component 1: LANA CITS Knowledge Base

The knowledge base consists of Tutorial Knowledge base, Arabic general contexts, user's profile and the log file. The Knowledge base design is draws on the knowledge engineering of the Arabic CA domain.

1. The Domain

The domain used to script the CA is concerned with teaching simple scientific topics such as the solar system to the targeted age group (10-16 years old). The scientific material is based on

educational resources provided to schools by the Ministry of Education in Saudi Arabia. Scripts consist of a context structured according to the topics in the science book such as Earth, Moon, Solar System, the eclipse, etc.). LANA CITS is therefore designed to deliver the Science topics to Autistic learners in Arabic language based on standard teaching in schools in Saudi Arabia.

2. Arabic CITS Knowledge Engineering

Base on Jee et al.[35] the definition of, knowledge engineering is the process of extracting, representing, encoding, and testing of expert knowledge. Knowledge engineering was used in LANA CITS to script the Arabic CA domain, which involved a number of process beginning with extraction, design, and development of the knowledge base. In the knowledge extraction process, a short interview was conducted with three primary school teachers in Saudi Arabia, who have experience in teaching the science subject to the targeted age group. The aim of this interview was to design and deliver the tutorials in LANA CITS as closely as possible to how the teachers deliver the subject in the classes. The information garnered from this interview was transcribed and used to create the tutorials. In the knowledge design process, the information from the interviews was then collated and used with the science book to design the tutorials. The designed tutorial was evaluated by the teachers to get a feedback. Once the tutorials were approved by the teachers, the tutorials were implemented in LANA CITS. In the knowledge implementation process, the knowledge base was developed in a relational MySQL database in order to store and retrieve the resources. The knowledge base for the LANA CITS consists of 2 main layers: the domain, which is the science tutorials layer and, a general contexts layer, illustrated in Figure 2.

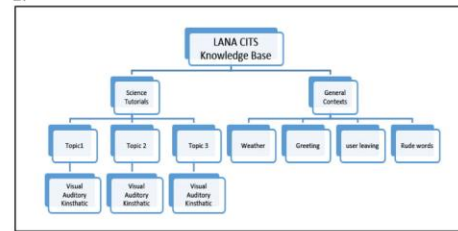


Figure 2: Knowledge base layers

Each layer represents a context, and each context has all the related sub contexts mapped to it. The domain layer holds all the tutorials sessions related to the science subject. The general contexts layer deals with general conversation not related to the domain, such as greeting and weather. The general context layer is included in to the knowledge base to make LANA CITS seem more intelligent by responding to the user's utterances, which are not related to the main domain. However, a few select sub-contexts have been implemented because it is not possible to cover all aspects of general contexts.

3. Mapping the tutorial into VAK learning style model

After the tutorial is designed and approved by the teachers then the tutorial questions are mapped to the VAK learning style

model. Each tutorial question is associated with the VAK learning style material. Each question in the tutorial is adapted and mapped to the Visual learning style through pictures and videos, the Auditory learning style through the addition of more audio and, the Kinesthetic learning style through the introduction of models and instructions.

4. Developing LANA CITS Scripting Language

In order to script the domain, the domain is organised into a number of contexts and each context contains rules, each rule in the domain contains a number of patterns and a response that forms the CA output to the user. The contexts represent the tutorial topics and the rules represent the agent's questions for such topic, while the patterns represent the user's utterance, which belong to such rule. Table 1 shows an example of the scripting language.

Context: 1
Context Name: Topic 1 (Earth, Sun, and Moon)
Rule Number: 1
Rule Name: Object moves in the gravitationally curved path called the orbit. The earth rotates around the sun and around its own axis. How long the earth takes to rotate around its axis?
Pattern: It * a day
Pattern: It * one day
Pattern: It * 24 hours
Pattern: * 24 hours
Response: Excellent, the earth takes one day (24 hours) to rotate around its axis
Image: yes
Audio: yes
Instruction: yes

Table 1: scripting language (translated)

B. Component 2: Designing and Implementing the Arabic CA

The Arabic CA architecture illustrated in Figure 1 was implemented. The components were developed using JAVA. The knowledge base and script databases were developed using the MySQL. The functions of the individual components of the Arabic CA architecture are described in detail in the following sections.

1. Controller

The controller is responsible for directing and managing the entire conversation. The controller works with several other components, which are the ITS manager, GUI and CA manager to achieve the conversation goal. First, the controller communicates with the ITS manager to know the student's learning style. Second, it gets the user's utterance and provides an utterance checking process. The controller will check the utterance based on the following constraints:

- Cleansing the utterance: The controller uses the utterance filter to remove special characters (i.e. \$, &, *, !, ?, ~, ^, (,), ^) from the utterance.
- Check for rude words: the system will warn the user three times before terminating the session.

After the utterance is parsed the controller then works together with the conversation agent manager to ensure that the conversation is within the tutorial scenario, or whether the user switched the context. After that the controller delivers the

response back to the user with the supporting material such as picture, audio, or instructions according to the user's learning style.

2. Conversational Agent Manager

The role of the Conversation Manager is to control the conversation to make sure that the goal is achieved, which is completing the tutorial. The conversation manager checks whether the user stays on the tutorial topic, or the user switch the context. Goal-oriented CAs should manage utterances in an intelligent way [36]. For example, when the CA manager receives an utterance that is not related to the tutorial topic, the CA manager checks the user utterance with the general contexts layer to see if the utterance matches other general contexts within the database. If the utterance matches one of these contexts, the CA manager generates the response to the user, and then the user is brought back to the point where the conversation is interrupted and then directed towards the goal.

3. Conversational Agent Engine

The LANA CA engine has two main components that were specifically developed to deal with the features unique to the Arabic language in terms of its grammatical and morphological nature. These components are: Pattern matching approach (Wildcard PM) and String similarity algorithm (Cosine algorithm). The combination of these components within LANA CA engine will help to calculate the similarity of the user utterance with scripted patterns using string similarity metrics. Consequently, reducing the need to cover all possible utterances when scripting the domain.

The first component of the Arabic CA engine, which is the primary phase of the engine, is pattern matching. It is based on similar Method to the Info Chat [23] of PM, where the user utterance is matched to stored patterns. These patterns contain wild card characters to represent any number of words of characters. For example, the wild card symbol (*) matches any number of words, where the wild card symbol (?) matches a single character. The second component of the Arabic CA engine is using the short text similarity algorithm, which is Cosine similarity [37]. It determines the similarity between two pieces of text by representing each piece of text in the form of word vector. A word vector is a vector of length N where N is the number of different tokens in the text. The similarity is computed as the angle between the word-vectors for the two sentences in vector space. For two texts t1 and t2 the cosine similarity is illustrated in the following equation:

$$SIM(t_1, t_2) = \frac{\sum_{i=1}^n t_{1i} t_{2i}}{\sqrt{\sum_{i=1}^n t_{1i}^2} \times \sqrt{\sum_{i=1}^n t_{2i}^2}} \quad (1)$$

The cosine similarity result is non negative and bounded between [0,1] where 1 indicates that the two text are identical.

4. LANA CA Engine Workflow

First the system will use PM (with wildcards) to match the user utterance with the pattern saved in the database, if a match is

not found, the next step is to use STS Cosine similarity. To illustrate the cosine similarity algorithm, assume the user utterance was (S1), while the pattern stored in the LANA database was (S2). As shown in Table 2:

User Utterance: S1 = ليس غلاف جوي باختلاف الأرض	S1: The moon does not have atmosphere whereas the earth has.
Stored pattern in LANA CITS: S2 = يختلف الغلاف الجوي عن الأرض بأن ليس له غلاف جوي	S2: The moon is different from the earth that it does not have an atmosphere.

Table 2: Cosine Similarity Example

The result obtained from Cosine similarity when applied to S1 and S2 is equal (0.95). The utterance is not recognised by the PM approach, whereas when the system applied the Cosine similarity algorithm, the utterance is recognised.

C. Component 3: LANA Intelligent Tutoring System (ITS)

This section describes the third component in LANA CITS architecture, which contains two main components, The Graphics User Interface, and The ITS manager.

1. Graphics User Interface

The GUI is the point where LANA CITS and the user interact with each other. LANA CITS has a character called LANA, which appears throughout the system interfaces in order to make it more engaging and natural for the users. Massaro and Bosseler [38] developed a computerized tutor called (Baladi) to teach vocabulary and functional language use to children with Autism with and without a cartoon face. They evaluated whether the face would increase the rate of learning for receptive measures. The results showed that the children's performance with the face condition was better than in the no face condition. Based on this, LANA's character was designed and evaluated by primary school teachers in order to make the tutorial more engaging. Figure 3 shows screenshot of the LANA GUI.



Figure 3: LANA GUI

2. ITS Manager

The ITS manager is responsible for personalising the tutorial according to the user's learning style, which is determined at the beginning of the tutorial through a questionnaire. The ITS manager can interact with: the conversation agent components through the controller, the Knowledge base, and the Graphics user interface. In order to personalise the tutorial session according to the student's learning style, a questionnaire will be applied first for each child. The questionnaire was developed on the basis of a widely disseminated version [39] to be completed by pupil with parents or teacher help. This questionnaire focused on Smith's visual, auditory, and kinaesthetic styles (VAK). There were three questions for each style and the pupils had to respond 'yes' or 'no' to each question. There are a total of nine questions in the questionnaire and the student's learning style result will be based on the highest score in one area.

IV. EXPERIMENTAL METHODOLOGY

In order to evaluate the implemented LANA CITS framework and its architecture, a pilot study will be conducted with Non-autistic children (age 10-16 years old). This pilot study aims to evaluate the effectiveness and robustness of LANA CITS in order to address any weakness and then to conduct the second study with Autistic children. To formulate evaluation metrics, the Goal Question Metric (GQM) methodology is used [40]. The (GQM) model is used to identify which metrics are required to gauge the effectiveness and robustness of LANA. A total of 24 participants will be involved with age group (10-16) years old and their first language is Arabic. The evaluation will be done through two experiments: The first experiment is to test two questions, which are:

- Question1: Can the learning Style VAK help the student to learn better?
- Question2: Can LANA CITS adapt the student's learning style?

The data will be gathered from the user feedback questionnaire and the pre- and post-test scores with and without using the VAK learning Style. The second experiment will be conducted to test two questions, which are:

- Question 1: Are the participants satisfied with the usability of LANA CITS?
- Question 2: is LANA CA robust?

The data will be gathered from the LANA log file and the user feedback questionnaire.

V. CONCLUSION AND FUTURE WORKS

This paper outlined a novel Arabic CITS called LANA, which used VAK learning style model to enhance the learning for children with Autism. LANA CITS has the main features:

- LANA CITS will be able to adapt the VAK learning style within the tutorial.
- LANA CITS will be able to converse in Arabic language to teach children age (10-16) the science subject.

- LANA engine will use PM with STS similarity algorithm to recognise the user utterance.

In future, a first pilot study will be conducted to collect preliminary data, using the experimental methods that were mentioned in the previous section. The pilot study aims to test the proposed architecture, and identify any details that need to be addressed before conducting the second pilot study with Autistic children on the target age group (10-16) years old whose first language is Arabic, who have high functioning autism (HFA) and have no problem with language and intellectual disabilities. The combination of the two testing stages will ensure the robustness and tutoring success of the final CITS.

VI. REFERENCES

- [1] C. Putnam and L. Chong, "Software and technologies designed for people with autism: what do users want?," presented at the 10th International ACM SIGACCESS conference on computers and accessibility, 2008.
- [2] D. Murray, *Autism and information technology: Therapy with computers*. (In Autism and Learning: A Guide to Good Practice). Brookes, 1997.
- [3] S. Parsons and S. Cobb, "State-of-the-art of virtual reality technologies for children on the autism spectrum," *European Journal of Special Needs Education*, vol. 26, no. 3, pp. 355-366, 2011.
- [4] O. Grynspan, J.-C. Martin, and J. Nadel, "Multimedia interfaces for users with high functioning autism: An empirical investigation," *International Journal of Human - Computer Studies*, vol. 66, no. 8, pp. 628-639, 2008.
- [5] D. Massaro, "Symbiotic Value of an Embodied Agent in Language Learning," presented at the Proceedings of the 37th Hawaii International Conference on System Sciences, 2004.
- [6] E. I. Limoges, L. Mottron, and C. B. Bolduc, Claude Godbout, Roger, "Atypical sleep architecture and the autism phenotype," *Brain*, vol. 128, 2005.
- [7] F. Baker. (2015). *Learning styles in children*. Available: <http://www.kidspot.com.au/school/primary/learning-and-behaviour/learning-styles-in-children>
- [8] V. Chislett. (2015). *VAK Learning Styles Self-Assessment Questionnaire*. Available: <http://www.businessballs.com/vaklearningstylestest.htm>
- [9] J. O'Shea, Z. Bandar, and K. Crockett, "Systems Engineering and Conversational Agents," 2011, vol. 10, pp. 201-232.
- [10] K. Crockett, O. S. James, and Z. Bandar, "Goal orientated conversational agents: applications to benefit society," in *Agent and Multi-Agent Systems: Technologies and Applications*: Springer, 2011, pp. 16-25.
- [11] A. Campbell, "TECHNOLOGY ASSISTED SEXUAL HEALTH AGENT (TASHA): A CONVERSATIONAL AGENT FOR PATIENT DECISION MAKING ON THE MANAGEMENT OF CERVICAL DYSPLASIA/CANCER," in *Making Difficult Clinical and Policy Decisions: The Example of Ageing and End of Life Care in Asia-Pacific*, 2016: Smdm.
- [12] S. Tegos, S. Demetriadis, and A. Karakostas, "Promoting academically productive talk with conversational agent interventions in collaborative learning settings," *Computers & Education*, vol. 87, pp. 309-325, 2015.
- [13] K. O'Shea, K. Crockett, and Z. Bandar, "Application of a semantic-based conversational agent to student debt management," in *Fuzzy Systems (FUZZ)*, 2010 IEEE International Conference on, 2010, pp. 1-7: IEEE.
- [14] M. Owda, Z. Bandar, and K. Crockett, "Information extraction for SQL query generation in the conversation-based interfaces to relational databases (C-BIRD)," in *Agent and Multi-Agent Systems: Technologies and Applications*: Springer, 2011, pp. 44-53.
- [15] O. G. Aloabaidi, K. A. Crockett, J. D. O'Shea, and T. M. Jarad, "Abdullah: An Intelligent Arabic Conversational Tutoring System for Modern Islamic Education," in *Proceedings of the World Congress on Engineering*, 2013, vol. 2.
- [16] J. Cassell, *Embodied conversational agents*. MIT press, 2000.
- [17] L. Yin, T. Bickmore, and D. E. Cortés, "The impact of linguistic and cultural congruity on persuasion by conversational agents," in *International Conference on Intelligent Virtual Agents*, 2010, pp. 343-349: Springer.
- [18] D. Crystal, "Front Matter," *A Dictionary of Linguistics and Phonetics*, Sixth Edition, pp. i-xxv, 2008.
- [19] S. Demir, "Improved treatment of word meaning in a Turkish conversational agent," Citeseer, 2003.
- [20] M. Fernezelyi, Z. S. Maszák, and R. Langh, "Smalltalk: interactive installation," in *Proceedings of the 14th ACM international conference on Multimedia*, 2006, pp. 1025-1026: ACM.
- [21] M. Hijawi, Z. Bandar, and K. Crockett, "The Enhanced Arabchat: An Arabic Conversational Agent," *International Journal of Advanced Computer Science and Applications*, 2016.
- [22] M. Kaleem, J. O'Shea, and K. Crockett, "Development of UMAIR the Urdu Conversational Agent for Customer Service," 2014.
- [23] D. Michie and C. Sammut, "Infochat scripser's manual," *Convagant Ltd., Manchester*, 2001.
- [24] R. S. Wallace, "ALICE: Artificial Intelligence Foundation Inc.," ed, 2008.
- [25] T. Bickmore and T. Giorgino, "Health dialog systems for patients and consumers," *Journal of biomedical informatics*, vol. 39, no. 5, pp. 556-571, 2006.
- [26] J. F. Allen, D. K. Byron, M. Dzikovska, G. Ferguson, L. Galescu, and A. Stent, "Toward conversational human-computer interaction," *AI magazine*, vol. 22, no. 4, p. 27, 2001.
- [27] M. Maragoudakis, B. Kladias, A. Tsopanoglou, K. Sgarbas, and N. Fakotakis, "Natural language in dialogue systems. A case study on a medical application," in *Proceedings of the Pan-Hellenic Conference, with International Participation, in Human-Computer Interaction, PCHCI, Patras Greece*, 2001, pp. 197-201.
- [28] V. Hatzivassiloglou, J. L. Klavans, and E. Eskin, "Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning," in *Proceedings of the 1999 joint sigdat conference on empirical methods in natural language processing and very large corpora*, 1999, pp. 203-212: Citeseer.
- [29] N. Y. Habash, "Introduction to Arabic natural language processing," *Synthesis Lectures on Human Language Technologies*, vol. 3, no. 1, pp. 1-187, 2010.
- [30] Y. Li, Z. Bandar, D. McLean, and J. O'Shea, "A Method for Measuring Sentence Similarity and its Application to Conversational Agents," in *FLAIRS Conference*, 2004, pp. 820-825.
- [31] J. O'Shea, Z. Bandar, K. Crockett, and D. McLean, "A comparative study of two short text semantic similarity measures," in *KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications*, 2008, pp. 172-181: Springer.
- [32] K. A. Papanikolaou, M. Grigoriadou, H. Kornilakis, and G. D. Magoulas, "Personalizing the Interaction in a Web-based Educational Hypermedia System: the case of INSPIRE," *User modeling and user-adapted interaction*, vol. 13, no. 3, pp. 213-267, 2003.
- [33] A. C. Graesser, P. Chipman, B. C. Haynes, and A. Olney, "AutoTutor: An intelligent tutoring system with mixed-initiative dialogue," *IEEE Transactions on Education*, vol. 48, no. 4, pp. 612-618, 2005.
- [34] A. Latham, K. A. Crockett, and Z. Bandar, "A Conversational Expert System Supporting Bullying and Harassment Policies," in *ICAART (I)*, 2010, pp. 163-168: Citeseer.
- [35] B. D. Jee et al., "Drawing on experience: how domain knowledge is reflected in sketches of scientific structures and processes," *Research in Science Education*, vol. 44, no. 6, pp. 859-883, 2014.
- [36] A. M. Latham, "Personalising learning with dynamic prediction and adaptation to learning styles in a conversational intelligent tutoring system," *Manchester Metropolitan University*, 2011.
- [37] G. Qian, S. Sural, Y. Gu, and S. Pramanik, "Similarity between Euclidean and cosine angle distance for nearest neighbor queries," in *Proceedings of the 2004 ACM symposium on Applied computing*, 2004, pp. 1232-1237: ACM.
- [38] D. W. Massaro and A. Bosseler, "Read my lips: The importance of the face in a computer-animated tutor for vocabulary learning by children with autism," *Autism*, vol. 10, no. 5, pp. 495-510, 2006.
- [39] A. Smith, *Accelerated learning in practice*. A&C Black, 1998.
- [40] N. E. Fenton and S. L. Pfleeger, "Software Metrics: A Rigorous and Practical Approach: Brooks," ed: Cole, 1998.

Can LANA CITS Support Learning in Autistic Children? A Case Study Evaluation

Sumayh S. Aljameel, James D. O'Shea, Keeley A. Crockett, Annabel Latham
Department of Computing, Mathematics and Digital Technology
Manchester Metropolitan University
sumayh.s.aljameel@stu.mmu.ac.uk, {j.d.oshea, k.crockett, a.latham}@mmu.ac.uk

Abstract—LANA CITS is a Conversational Intelligent Tutoring System that uses the Visual, Auditory, and Kinaesthetic learning style (VAK). It supports learning in autistic pupils, who are studying in mainstream primary schools. Facilitating the learning of these pupils using traditional teaching within mainstream schools is complex and poorly understood. This paper presents investigation into how LANA CITS using VAK learning style model can be adapted to autistic pupils learning style and improve their learning in mainstream schools. This paper provides a case study evaluation of three children with high-functioning autism examining the effectiveness of learning with LANA CITS. The case study took place in primary school in Saudi Arabia. The results were positive with the students engaged in the tutorial and the teacher noticed some improvement over classroom activities. This results support for the continuing development, evaluation, and use of CITS for pupils with autism in mainstream schools.

Keywords—Intelligent tutoring system; autism spectrum disorder; conversational agent; VAK learning style

I. INTRODUCTION

Pupils with Autism Spectrum Disorder (ASD) are affected in different degrees in terms with their level of intellectual ability. Some people with Asperger syndrome or high functioning autism are very intelligent academically but they still have difficulties in social and communication skills [1]. In recent years, many of these pupils are taught within mainstream schools. However, the process of facilitating their learning and participation remains a complex and poorly understood area of education [2], [3]. Many teachers in mainstream schools encourage the principles of inclusive education, but they do not feel that they have the enough training to teach autistic pupils [4], [5]. Many pupils with ASD tend to be affected differently as a result of the same symptoms, so it is important to know how each child is impacted on in terms of their cognitive and interaction abilities. Hence, the specific requirements of each child, which is known as learning style, should be identified [6]. Pupils with ASD are more likely to rely on one style of learning, which could be determined by observing the person [7]. VAK theory is considered to be one of the classical learning theories in the educational field [8]. It classifies learners by sensory preferences: Visually (V), Auditory (A), and Kinaesthetic (K) [9]. Many researches [10]–[12] found that many people with autistic spectrum are visual learners as they can learn well through watching videos, pictures and movies. Another study [13] found that Autistic people have learning preferences of

kinaesthetic, visual, then auditory. Roberts [14] found that people with autism who are kinaesthetic learners, learn better when include “hands-on” activities with the teacher taught material to achieve learning. Nevertheless, research that has been conducted with autistic children regarding the impact of learning style is limited and need further research.

Technologies have attracted increasing attention in the autism community for their educational [15]–[17]. Several studies [18], [19] reported that the majority of people with autism feel positive towards technology and computer-based training. This is attributable to the fact that software programs offer a structured and predictable environment, which accommodates the autistic children preferences for routine and repetitive behaviours [20]. Truong [21] reviewed 51 studies of the recent applications which examined the integration of learning styles theories in adaptive learning system. In this review, it reported a positive impact of adaptive systems using learning styles on students' satisfaction. No known intelligent tutoring system using the VAK learning style model has been specific developed for autistic children.

This paper is a part of our investigation within LANA CITS project [22], [23]. This project examines the research and development of a novel Arabic conversational intelligent tutoring system (CITS) for children with autism spectrum disorder (ASD) using the Visual, Auditory and Kinaesthetic learning styles model (VAK). The project aims to design and implement an Arabic CITS for children with autism aged 10–16 using the (VAK) model. The CITS architecture used the pattern matching (PM) and the Arabic short text similarity (STS) to extract the responses from resources in particular domain (Science subjects). This paper presents a case study evaluation of LANA CITS used by three pupils with high-functioning autism or Asperger's, who are studying in mainstream schools in Saudi Arabia. The case study approach was chosen for this study because characteristics of pupils with ASD differ widely from each other. In addition, large groups of pupils with ASD who have similar characteristics are difficult to recruit. The case study examines the effectiveness of using LANA CITS and whether the autistic pupils satisfied with the usability of LANA CITS. The main contribution of this paper is that it is the first case study evaluation of CITS using VAK model for pupils with high functioning autism in mainstream schools. This paper is organised as follows: Section II provides a brief overview of LANA CITS. Section III describes the case study

methodology. Section IV presents the results and discussion and Section V, the conclusion.

II. LANA CONVERSATIONAL INTELLIGENT TUTORING SYSTEM

A. Background

Conversational Intelligent Tutoring Systems (CITS) provide a natural language interface to an online tutoring system to allow users to learn topics through discussion as they would in the classroom. Many researchers have developed CITS for different domains, such as AutoTutor [24] which is a computer tutor that is designed to assist students within university to learn an introductory computer literacy course. However, AutoTutor does not consider the student learning styles during a tutoring session. Oscar [25] is another example of CITS, which overcomes AutoTutors limitation by predicting and adapting dynamically to a student's learning style. Oscar is developed for students in University's level and teaches SQL course. In the Arabic language, CITS is considered as a new area of research. Abdullah CITS [26] is a recent of Arabic CITS for modern Islamic education. Abdullah CITS is an online system that aims to teach children aged 10 to 12 years old essential topics in Islam. Abdullah CITS does not consider the student learning styles. To our knowledge, no academic research exists on the Arabic conversational intelligent tutoring system for Autistic children.

B. LANA CITS

LANA CITS [22], [23] is a novel Arabic CITS, which delivers topics related to the science subject by engaging with the user in Arabic language. LANA uses Visual, Auditory, and Kinaesthetic (VAK) learning style model to personalise the lesson according to the user learning style preference. The main features of LANA CITS are:

- Ability to control the conversation through context. The domain used to script the CA was from the Science curriculum for the age (10-12 years old). The material was provided by the Ministry of Education in Saudi Arabia [27]. Scripts consisting of a context structured according to the topics in the science book such as the Earth, Moon, Solar System, eclipse, etc.).
- Ability to personalise the lesson with the user's learning style (VAK) and provide suitable material to the user according to the user's learning style (images and videos for Visual learner, Sound for Auditory learner, physical models for Kinaesthetic learner).
- A conversation agent scripting language, which provides Arabic dialogue for LANA CITS allowing capturing variables from the user utterance.
- A novel conversation agent engine using the pattern matching (PM) technique was first used to find a suitable response to the user's utterance. If it is not found, the engine uses the Short Text Similarity (STS) technique to find a response. Combining the techniques is beneficial to reduce the number of scripted patterns.

Fig. 1 shows a screenshot of the conversation interface of LANA CITS, where the system converses with the student by asking him/her a question about a specific science topic and then providing the suitable response based on the student's utterance.



Fig. 1. LANA CITS graphics user interfaces.

III. CASE STUDY

This section presents the case study, which took place in a primary school in Saudi Arabia. Ethical informed consent was obtained from all participants, parents and the School. The study aims to test hypothesis H1 (LANA CITS can support learning in autistic children).

A. Participants

In this study, three participants who met the following criteria were interviewed and observed:

- Participant with age group (10-12) years old.
- He/she had previously received a community clinical diagnosis of high functioning autism spectrum disorder or Asperger's, indicating that they have no problem with language and intellectual impairments.
- His/her first language was Arabic.
- Studying in mainstream primary schools in Saudi Arabia.
- No previous experience using LANA CITS.

To ensure data privacy, the names of participants have been changed in this paper. In this study, three autistic students are involved and all of them are female (Mariam, Fatimah, and Dana). Two of them (Mariam and Fatimah) are in (Grade 5), and their age is 11 years old, the third students (Dana) is in (Grade 4) and her age is 10 years old. The three participants are studying in mainstream primary school, which provided educational provision for children with ASD in Saudi Arabia. All three participants have Arabic as their first language and none of them had a previous experience using LANA CITS. Each child had previously received a community clinical diagnosis of Asperger's, who they have no problem with language and intellectual impairments.

B. Ethical Considerations

Because the participants were children, an informed consent form and information sheet were sent to the participant's parents and the School. The information sheet outlined the details and purpose of the study, and the benefits for participating in the study. They also informed that the information gathered from the participant is stored securely and none of children's real names will be recorded.

C. Data Collection Methods

There are several data collections methods in qualitative research. Marshall and Rossman [28] identified three primary methods that qualitative researchers use for gathering information. Primary methods involve gathering information from the source directly. These three methods are: 1) participating in the setting, 2) observing directly, and 3) interviewing in depth. In this study, two primary methods are used: Observation and Interviewing. "Observation involves the systematic noting and recording of events, behaviors, and artifacts in the social setting," [28]. Kvale and Brinkman [29] defined the interview as a conversation with structure and a purpose. There are secondary methods that may also be used to gather information. This involves gathering information from someone or something else outside of the original source by analysing documents such as: videos, photography, interaction analysis, and questionnaires [28].

1) Interviewing

Interviewing is a universal methods in qualitative research [30], [31]. Interviews are typically divided into three categories: structured, unstructured, and semi-structured [32]. In structured interviews, all participants will have the same questions using the same wording and in the same order [33]. However, while a structured interview ensures that comparable responses are gathered from each participant, it leaves no room for elaboration [34]. The second category is an unstructured interview, which contains open questions and subsequent questions depending on the participant's responses [35]. The drawbacks of this type of interview are the difficulty of processing data and the effort required. It is difficult to find similar statements from different participants and hence the links are often difficult to make. The most common type of interviews used in qualitative research are the third category, which is semi-structured interviews. It involves structured questions and the researcher is free vary the order and wording of the questions [32], depending on the direction of the interview, and to ask additional questions. According to Berg [34], the researcher is able to adjust the wording of the questions, and add or delete probes, as necessary. With this structure, the interviewer guided the interview, to obtain a good description from the participant's perspective. Moreover, this structure allowed the researcher the flexibility to dive deeper when appropriate. Creswell [36] asserted that the semi-structured process lets the researcher to explore and ask follow-up questions if necessary to clarify the particular subject. This was an extremely beneficial process in the field of qualitative interviewing.

The Semi-structured interviewing was used in this study because the participants are children and the level of

comprehension may vary. The interview was interactive and the participants were able to ask questions. In addition, notes were taken during the interview to capture the core of the participant's comments and to note any non-verbal behaviour and facial expressions. In this study, two interviews were conducted for each student, the first interview was with the autistic student and the second interview was with the teacher. The first interview, which was with the autistic student, contains of nine questions and the interviewer was free to ask follow-up questions. Table I shows the nine questions that are used in the interview. The questions were asked in Arabic.

TABLE I. THE INTERVIEW SHEET WITH THE AUTISTIC STUDENT

1. Do you think LANA Tutor is enjoyable to use?
• Which part do you like?
• Why do you like this part?
2. Do you think LANA had a right amount of information in one Tutorial?
3. Do you remember what LANA has taught you in the tutorial?
• Tell me in one or two things you learned from LANA
4. Did you like using your learning style (Visual – Auditory - Kinaesthetic) in LANA?
• Would you like to use your learning style with all subjects and tutorials?
5. Do you like the way LANA looked?
• What do you think of the color of LANA?
• Any suggestion to add in the design or the character.
6. Do you think you can learn from LANA better than your teacher?
7. Would you like to use LANA again with other lessons?
8. Is talking with LANA like talking with your friend?
9. Would you like to use LANA in the classroom?

The second interview was conducted with the teacher. The aim of this interview is to measure the student's interaction and behaviour from the teacher's point of view. The interview contains five questions as shown in Table II.

TABLE II. THE INTERVIEW SHEET WITH THE TEACHER

1. Is there any behavioural issue with the student?
2. What is the student's education level in the school?
3. How was the student's interaction during the tutorial?
4. Did the student interact with the LANA's tutorial in the same way as the class tutorial?
5. Do you have any suggestions to add to LANA?

2) Participant Observation

In addition to semi-structured interviews, the participant has been observed during the interview by Sumayh, the lead researcher of this study. Observations allow the interviewer to observe the context in which the participants interact [36]. Additionally, observations allow the researcher to notice the participant's behaviors, facial expressions, and the interactions.

D. Data Analysis

Following the interviews the data was analysed using the following methodology [36].

1) Data management: This involves organizing and gathering all the data. In this study, the interview was transferred into a written format.

2) Reading and noting: This involves reading through all the transcriptions and making notes in the margins regarding the interviewee's comments.

3) Describing: This involves describing the participant's experience with the study.

4) Classifying: The researcher then makes a list of significant statements, quotes that emerged from the data and then grouped together.

5) Interpreting: the researcher develops textual and structural descriptions of the data. Textual descriptions are the written descriptions of what the participants experienced. After the textual descriptions are written, the researcher outlined the contextual influence, and this is the structural description.

6) Representing and visualizing: After both the textual and structural descriptions are written, the researcher writes the findings of the study. In addition, tables will be created to help illustrate the data.

E. Participant's Interaction

The study started by detecting the participant's learning style using a child friendly-customised VAK learning style preferences questionnaire, which was answered by the student with the help of the teacher. The LANA CITS registration interface was then shown to the participant and the lead researcher registered the participant with her learning style into the system in order to adapt the tutorial according to her learning style. Then, a pre-test interface started (see Fig. 2) and the participant was asked to complete the pre-test by choosing the right answer using the touch screen. When the pre-test was completed, the tutorial started and the participant interacted with the tutorial and answered the questions related to the science subject. During the tutorial, materials were used according to the learning style (Visual: video and images, Auditory: sounds, Kinaesthetic: physical models). After the participant completed the tutorial, a post-test interface started and the participant answered the questions in the post-test. When the tutorial was completed a verbal interview (see Table I) was conducted in order to measure the participant satisfaction and their opinion about using LANA CITS. During the tutorial and the interview, the student was observed by the lead researcher and the teacher to notice the participant's behaviors, facial expressions, and the interactions. Support was provided to the participant if asked. When the student finished and left the room, the teacher was interviewed by the lead researcher (see Table II).



Fig. 2. Pre-test Questions interface.

IV. RESULTS AND DISCUSSION

In this section, the data that was gathered from the interviews (Tables I and II) and the observation were described and classified in order to test Hypothesis H1 (LANA CITS can support learning in autistic children).

A. First Participant Results and Discussion

Mariam: The first participant interviewed in this study, was Mariam. She is eleven years old in Grade 5 and she has an early community clinical diagnosis of Asperger's, with no problem with language and intellectual impairments. Based on the teacher's interview, Mariam made an improvement in her behaviour during her study in the school from Grade 1 to 5. Her education level is very good especially in Maths and English. She is interested in using technology such as a computer, iPad, and mobile devices. The study begun by detecting Mariam's learning style using the questionnaire. The results from the questionnaire showed that her learning style was Visual. Mariam completed the pre-test and then interacted with the tutorial and answered the questions. During the tutorial, videos/images appeared for each information in the tutorial (see Fig. 3). After she completed the tutorial, a post-test interface started and she answered the questions in the post-test. Mariam enjoyed using LANA CITS as indicated in her interview answers. She said that she especially liked watching the videos and the pictures during the tutorial. She was able to use LANA CITS independently with ease and required no additional support once told in how to use it. Her learning gain was improved when comparing the pre-test and the post-test scores. Table III shows the pre-test and the post test scores for all three participants. She was able to remember what she had learned from LANA, when she was asked in the interview. The rest of the interview questions were positive but when she was asked if she like learning from LANA CITS more than learning from her teacher, she answered 'Maybe'.



Fig. 3. The tutorial screenshot for the visual learner.

B. Second Participant Results and Discussion

Fatimah: The second participant interviewed in this study was Fatimah. She is eleven years old in Grade 5 and she had a late diagnosis of Asperger's, when she was nine years old. Based on the teacher's interview, Fatimah's education level is below the average because she does not like to study although she has an excellent memory. She has the ability to remember everything she has seen or read. Her learning style was detected using the questionnaire and it was found that her learning style was Visual. Fatimah completed the pre-test and then interacted with the tutorial and answered the questions. When she completed the tutorial, a post-test interface started and she answered the questions in the test. Fatimah said that using LANA CITS was very enjoyable. She liked watching to the videos and the pictures during the tutorial. She was using LANA CITS independently and easily with no additional support once she was told in how to use it. According to her scores in pre-test and post-test (Table III), her learning gain was improved the post-test score was grater that the pre-test score. She was able to remember what she has been learned from LANA, when she was asked in the interview. When she was asking if she like learning from LANA CITS more than learning from her teacher, she answered 'Both'. The rest of the questions were answered positively.

C. Third Participant Results and Discussion

Dana: The third participant interviewed in this study was Dana. She is ten years old in Grade 4 and she also had a late diagnosis of Asperger's, when she was nine years old. Based on the teacher's interview, Dana's education level is above the average. She is very good student in all subjects. Her learning style was detected using the questionnaire and it was found that her learning style was Kinaesthetic so she used physical models within the tutorial. Dana completed the pre-test and then interacted with the tutorial. Fig. 4 shows how the kinaesthetic learner interacts with the tutorial using physical model. When she completed the tutorial, a post-test interface started and she answered the questions in the test.



Fig. 4. The kinaesthetic learner interacts with the tutorial using physical model.

Dana was very interested when using LANA CITS and during the tutorial she was saying comments with positive words such as 'nice', 'I like it', 'I know the answer'. She was very excited when she used the magnetism model and the electric circuit during the tutorial. She was using LANA CITS independently and easily with no additional support once she was told in how to use it. She made a good improvement on her learning gain based on her scores in pre-test and post-test (Table III). The post-test score was greater than the pre-test score. When she was asked in the interview if she can remember the tutorial, she said that she remembered it and recalled information about the tutorial. She said that she preferred to learn from LANA more than from her teacher. The rest of the questions were answered positively.

TABLE III. THE PRE-TEST AND POST-TEST SCORES FOR THE THREE PARTICIPANTS

Student's Name	Pre-Test (out of 10)	Post-Test (out of 10)
Mariam	4	9
Fatimah	5	10
Dana	6	10

D. Teacher Observations and Suggestions

During the experiment, the teacher was attending in order to observe the student and to note if there is any different in the student's interaction from the classroom teaching. The teacher reported that Mariam interacted positively with LANA CITS and her interaction with LANA was better than in the classroom. According to the teacher's comments, Mariam does not like to listen to the lesson in the classroom for more than 15 minutes, she feels bored and she starts to do another task. Whereas, during the experiment, Mariam spent 25 minutes and she completed the whole tutorial. With the second participant, the teacher said that Fatimah interacted with LANA better than in the classroom. Based on the teacher's comments, Fatimah feels sleepy in the classroom and sometimes she sleeps during the lessons, whereas in the

experiment, Fatimah spent 30 minutes, she did not feel sleepy and she completed the whole tutorial. With the third participant, the teacher reported that Dana also interacted with LANA better than in the classroom because Dana likes learning from the computer more than learning from the teacher. The teacher indicated that the use of LANA CITS was beneficial for the autistic students. The teacher particularly liked idea of the using the student's learning style in the tutorial and this will meet the individual learning needs for the students with ASD. The teacher suggested that using LANA in the classroom would be beneficial because it will increase the students' independence and decreased their dependence on the teacher. For example, students with using LANA CITS will not need to wait for a teacher to explain the lesson and then the teacher can focus on other aspects such as the student's behaviour, and social skills.

V. CONCLUSION AND FUTURE WORKS

This paper outlined and detailed the case study that was used to evaluate the LANA CITS with three autistic students in a mainstream school in Saudi Arabia. The results from this study indicated that the three participants showed an improvement on their learning gain when comparing the scores of the pre-test and the post-test (Table III). Overall, a few differences were noted in terms of their perception of learning from LANA more than learning from the teacher. One student said that she liked the learning from LANA more than the learning from the teacher and the second student answered 'Maybe' and the third said 'Both'. Based on the teacher observation, LANA CITS was beneficial for the autistic students and the student's interaction during the tutorial was more positive than their interaction in the classroom. The teacher reported that one student feels sleepy and the other feels boring all the time in the classroom, whereas when they used LANA they were very interested and excited, because many students with ASD enjoy the use of technology. Finally, the teacher indicated that using LANA in the classroom would be beneficial because it will increase the students' independence and decreased their dependence on the teacher. Based on these results, it can be concluded that there is enough evidence to support H1 (LANA CITS can support learning in autistic children).

Future work will need with larger number of autistic students in mainstream schools in order to provide a quantitative estimate of the effect of LANA CITS.

REFERENCES

- [1] N. Bauminger-Ziveli, S. Eden, M. Zancanaro, P. L. Weiss, and E. Gal, "Increasing social engagement in children with high-functioning autism spectrum disorder using collaborative technologies in the school environment," *Autism*, vol. 17, no. 3, pp. 317-339, 2013.
- [2] J. Barnard, A. Prior, and D. Potter, "Autism and inclusion: Is it working," London: National Autistic Society, 2000.
- [3] P. Davis, L. Florian, and M. Ainscow, Teaching strategies and approaches for pupils with special educational needs: A scoping study. DfES Publications, 2004.
- [4] E. M. C. Gregor and E. Campbell, "The attitudes of teachers in Scotland to the integration of children with autism into mainstream schools," *Autism*, vol. 5, no. 2, pp. 189-207, 2001.
- [5] K. Robertson, B. Chamberlain, and C. Kasari, "General education teachers' relationships with included students with autism," *Journal of Autism and developmental disorders*, vol. 33, no. 2, pp. 123-130, 2003.
- [6] L. D'Elia, G. Valeri, F. Sonnino, I. Fontana, A. Mammone, and S. Vicari, "A longitudinal study of the TEACCH program in different settings: The potential benefits of low intensity intervention in preschool children with autism spectrum disorder," *Journal of autism and developmental disorders*, vol. 44, no. 3, pp. 615-626, 2014.
- [7] S. M. Edelson, "Learning styles and autism," Center for the Study of Autism, Salem, Oregon. En <http://www.autism.org/styles.html>, 2000.
- [8] Mackay, Motivation, ability and confidence building in people. Routledge, 2010.
- [9] P. Miller, "Learning Styles: The Multimedia of the Mind. Research Report," 2001.
- [10] Banire, N. Jomhari, and R. Ahmad, "Visual Hybrid Development Learning System (VHDLs) Framework for Children with Autism," *Journal of Autism and Developmental Disorders*, vol. 45, no. 10, pp. 3069-3084, 2015.
- [11] O'Connell et al., Supporting Children with Autistic Spectrum Disorders. Routledge, 2013.
- [12] K. Snyder, T. S. Higbee, and E. Dayton, "PRELIMINARY INVESTIGATION OF A VIDEO-BASED STIMULUS PREFERENCE ASSESSMENT," *Journal of Applied Behavior Analysis*, vol. 45, no. 2, pp. 413-418, 2012.
- [13] M. Lisle, "Assessing learning styles of adults with intellectual difficulties," *Journal of Intellectual Disabilities*, vol. 11, no. 1, pp. 23-45, 2007.
- [14] K. D. Roberts, "Topic areas to consider when planning transition from high school to postsecondary education for students with autism spectrum disorders," *Focus on Autism and Other Developmental Disabilities*, vol. 25, no. 3, pp. 158-162, 2010.
- [15] M. L. Diener, C. A. Wright, S. D. Wright, and L. L. Anderson, "Tapping into technical talent: using technology to facilitate personal, social, and vocational skills in youth with autism spectrum disorder (ASD)," in *Technology and the treatment of children with autism Spectrum disorder*: Springer, 2016, pp. 97-112.
- [16] S. Panzavolta, "Virtual Reality as a Tool for Enhancing Learning in At-Risk Students and Increasing School Inclusion," in *Handbook of Research on Collaborative Teaching Practice in Virtual Learning Environments*: IGI Global, 2017, pp. 242-257.
- [17] S. S. Aljameel, J. D. O'Shea, K. A. Crockett, and A. Latham, "A Review of Current Technology-Based Intervention for School Aged Children with Autism Spectrum Disorder," in *Proceedings of SAI Intelligent Systems Conference*, 2016, pp. 868-879: Springer.
- [18] S. Bernardini, K. Porayska-Pomsta, and T. J. Smith, "ECHOES: An intelligent serious game for fostering social communication in children with autism," *INFORMATION SCIENCES*, vol. 264, pp. 41-60, 2014.
- [19] G. Lorenzo, A. Lledó, J. Pomares, and R. Roig, "Design and application of an immersive virtual reality system to enhance emotional skills for children with autism spectrum disorders," *Computers & Education*, vol. 98, pp. 192-205, 2016.
- [20] K. Gillespie-Lynch, P. J. Brooks, C. Shane-Simpson, N. L. Gaggi, D. Sturm, and B. O. Ploog, "Selecting Computer-Mediated Interventions to Support the Social and Emotional Development of Individuals with Autism Spectrum Disorder," *Special and Gifted Education: Concepts, Methodologies, Tools, and Applications: Concepts, Methodologies, Tools, and Applications*, p. 32, 2016.
- [21] H. M. Truong, "Integrating learning styles and adaptive e-learning system: Current developments, problems and opportunities," *Computers in Human Behavior*, vol. 55, pp. 1185-1193, 2016.
- [22] S. Aljameel, J. O'Shea, K. Crockett, A. Latham, and M. Kaleem, "LANA: An Arabic Conversational Intelligent Tutoring System for Children with ASD," ed: IEEE Transaction on Learning Technologies, 2018.
- [23] S. S. Aljameel, J. D. O'Shea, K. A. Crockett, A. Latham, and M. Kaleem, "Development of An Arabic Conversational Intelligent Tutoring System for Education of Children with ASD," in *Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, 2017 IEEE International Conference on, 2017, pp. 24-29: IEEE.

- [24] C. Graesser, P. Chipman, B. C. Haynes, and A. Olney, "AutoTutor: An intelligent tutoring system with mixed-initiative dialogue," *IEEE Transactions on Education*, vol. 48, no. 4, pp. 612-618, 2005.
- [25] M. Latham, K. A. Crockett, D. A. McLean, B. Edmonds, and K. O'Shea, "Oscar: An intelligent conversational agent tutor to estimate learning styles," in *Fuzzy Systems (FUZZ)*, 2010 IEEE International Conference on, 2010, pp. 1-8: IEEE.
- [26] O. G. Alobaidi, K. A. Crockett, J. D. O'Shea, and T. M. Jarad, "Abdullah: An Intelligent Arabic Conversational Tutoring System for Modern Islamic Education," in *Proceedings of the World Congress on Engineering*, 2013, vol. 2.
- [27] M. o. Education. (2017). Ministry of Education. Available: <https://www.moc.gov.sa/en/Pages/default.aspx>
- [28] Marshall and G. Rossman, "The how of the study: Building the research design," *Designing qualitative research*, pp. 55-101, 2006.
- [29] S. Kvale and S. Brinkmann, "Learning the craft of qualitative research interviewing," *Thousands Oaks: Sage Publications*, 2009.
- [30] Byrne, "Qualitative interviewing," *Researching society and culture*, vol. 2, pp. 179-192, 2004.
- [31] M. Q. Patton, *Qualitative research*. Wiley Online Library, 2005.
- [32] S. Q. Qu and J. Dumay, "The qualitative research interview," *Qualitative research in accounting & management*, vol. 8, no. 3, pp. 238-264, 2011.
- [33] B. Kajornboon, "Using interviews as research instruments," *E-journal for Research Teachers*, vol. 2, no. 1, pp. 1-9, 2005.
- [34] L. Berg. *Methods for the social sciences*. Pearson Education Inc, United States of America, 2004.
- [35] O. Doody and M. Noonan, "Preparing and conducting interviews to collect data," *Nurse researcher*, vol. 20, no. 5, pp. 28-32, 2013.
- [36] J. W. Creswell and J. D. Creswell, *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications, 2017.

LANA-I: An Arabic Conversational Intelligent Tutoring System for Children with ASD

Sumayh Aljameel, James O'Shea, Keeley Crockett, Annabel Latham, Mohammad Kaleem

Department of Computing, Math and Digital Technology

Manchester Metropolitan University

Sumayh.s.aljameel@stu.mmu.ac.uk, {j.d.oshea, k.crockett, a.latham, m.kaleem}@mmu.ac.uk

Abstract. Children with Autism Spectrum Disorder (ASD) share certain difficulties but being autistic will affect them in different ways in terms of their level of intellectual ability. Children with high functioning autism or Asperger syndrome are very intelligent academically but they still have difficulties in social and communication skills. Many of these children are taught within mainstream schools but there is a shortage of specialised teachers to deal with their specific needs. One solution is to use a virtual tutor to supplement the education of children with ASD in mainstream schools. This paper describes research to develop a novel Arabic Conversational Intelligent Tutoring System, called LANA-I, for children with ASD that adapts to the Visual, Auditory and Kinaesthetic learning styles model (VAK) to enhance learning. This paper also proposes an evaluation methodology and describes an experimental evaluation of LANA-I. The evaluation was conducted with neurotypical children and indicated promising results with a statistically significant difference between user's scores with and without adapting to learning style. Moreover, the results show that LANA-I is effective as an Arabic Conversational Agent (CA) with the majority of conversations leading to the goal of completing the tutorial and the majority of the correct responses (89%).

Keywords: Autism Spectrum Disorder, Conversational Agent, Intelligent Tutoring System, String Similarity, The Arabic Language.

1 Introduction

The number of children being diagnosed with autism spectrum disorder (ASD) is increasing [1]. Children with high functioning Autism (HFA) or Asperger's Syndrome (AS) (i.e. those with higher verbal IQ) are usually offered education in the mainstream schools. However, many mainstream schools are not able to include students with ASD because of the lack of skilled teachers and the poor training and provisions from the responsible institutions [2]. In addition, traditional education using human tutors is a challenge for students with autism, who have difficulties in communication and social interactions. Many researchers have reported that using a virtual tutor with students with ASD could meet the individual students needs [3, 4].

Conversational Intelligent Tutoring System (CITS) integrates natural language interfaces to allow users to learn topics through discussion as they would in the classroom. Many researchers have developed CITS for different domains. To our knowledge, no academic research exists on Arabic CITS developed specifically for Autistic children. LANA-I [5] is an Arabic CITS, which engages autistic children with a science tutorial where the curriculum material is mapped to VAK model. One challenge in building such a system is the requirement to deal with the Arabic language grammatical features and its morphological nature. The research into Conversational Agent (CA) development techniques revealed that hybrid approach was the most appropriate approach to develop an Arabic CA [6]. The engine of LANA-I is based on the two main CA development strategies, A Pattern Matching (PM) engine and a Short Text Similarity (STS) algorithm that calculate the matching strength of a pattern to the user utterance. The two parts of the engine work together to overcome some of the unique challenges of the Arabic language and to extract responses from resources in a particular domain (Science topic). The main contributions in this paper are:

- A novel architecture for an Arabic CITS using VAK model for an appropriate education scenario.
- The results of designing an experimental methodology to validate the educational tutoring scenario in the Arabic CITS.

In order to evaluate LANA-I, an initial pilot study was conducted on the general population. This study took place in UK with neurotypical children from the target age group (10-12) years whose first language is Arabic. It is important to test LANA-I, with the general population before testing it with Autistic children to avoid any problems and issues that may occur in the tutorial or confusion in the presentation of the tutorial material. The study used the learning gain measurement (defined in section 3) to evaluate the ability of CITS to adapt to a child's learning styles.

This paper is organized as follows: Section 2 describes the architecture, and methodology of implementing LANA-I. Section 3 explains the experimental methodology and the experiments. Section 4 provides the results and discussion. Finally, section 5 presents the conclusions and future works.

2 LANA-I Architecture and Implementation

LANA-I was developed based on two phases. The first phase involved designing and implementing an Arabic CA, and the second phase focused on development of an educational tutorial on science, mapping the tutorial to the VAK model and introducing the Arabic ITS interface to the CA. In the first phase, a new architecture for developing an Arabic CA was developed using both PM and STS. The key features are summarised as follows:

- Ability to control the conversation through context.
- Ability to personalise the lesson with the user's learning style (VAK) and provide suitable material to the user according to their learning style (images and videos for Visual learner, Sound for Auditory learner, physical models for Kinaesthetic learner).

- A CA scripting language, which provides Arabic dialogue for LANA-I CITS allowing capturing variables from the user utterance.
- A novel CA engine using the PM technique first to find a suitable response to the user's utterance. If it is not found, the engine uses the STS technique to find a response. Combining the techniques is beneficial to reduce the number of scripted patterns.
- Managing the response when the context is changed. For example, creating the right response when the user writes something that is not related to the tutorial topic.

The proposed framework for the Arabic CA consists of five components as shown in Figure1:

1. Graphical User Interface: Responsible for the communication between the user and CA (in this case the CITS tutor) through a web interface with panels to display supporting material.
2. Controller: The controller manages the conversation between the user and the Arabic CA, as well as cleaning the user utterance by removing diacritics and other illegal characters (e.g. ! £ \$).
3. Conversation Agent Manager: Responsible for controlling and directing the conversation through contexts. In addition, the manager ensures that the discussion is directed towards completion of the tutorial.
4. Conversation Agent engine: Responsible for pattern matching and calculating the similarity strength between the user utterance and the scripted patterns.
5. The knowledge base: The knowledge base is responsible for holding the tutorial domain knowledge in a relational database, which includes CA scripts, Log file, and General contexts such as weather, agreement and rude words.

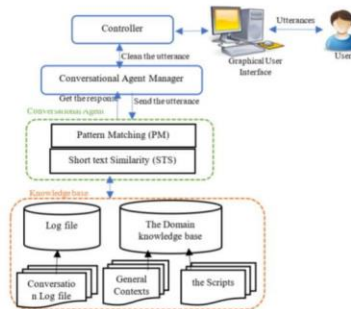


Fig. 1. The LANA-I CA architecture

In the second phase the ITS architecture is designed to adapt to the user's learning style within the Arabic CA. Based on the typical ITS architecture [7], LANA-I ITS architecture consists of four main components as shown in Figure 2:

1. User Interface Model: responsible for the interaction between the user and the ITS components.
2. Tutor Model: the ITS manager, which is the main component that interacts with the user through the GUI, and personalises the tutorial based on the user's learning style.
3. Student Model: a temporal memory structure, which records the user's responses during the tutoring session and the student's profile such as user ID, user's age, gender, user's learning style, and Pre-test and Post-test scores.
4. Domain Model: The Tutorial Knowledge Base, which contains structured topics that are presented to the user.

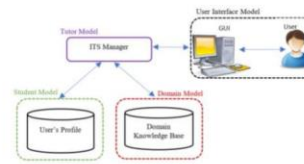


Fig. 2. LANA-I ITS Architecture

The LANA-I architecture combines the Arabic CA (Figure 1) and ITS architecture (Figure 2). The proposed LANA-I CITS architecture, shown in Figure3, contains three main components that are described in the following sections:

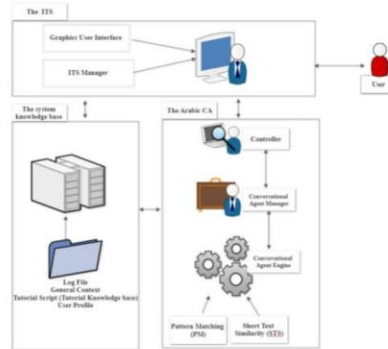


Fig. 3. LANA-I CITS architecture

2.1 LANA-I Knowledge Base

The knowledge base consists of four sub-components: (1) the Tutorial Knowledge base, (2) Arabic general contexts (e.g. weather, and greetings), (3) user's profile, and (4) the log file. The tutorial domain CA was from the Science curriculum for ages 10-

12. The material was provided by the Minis-try of Education in Saudi Arabia [8]. Script contexts were structured according to the topics in the science book taught in schools in Saudi Arabia, such as the Earth, Moon, Solar System, eclipse, etc. To design the knowledge base, a short interview was conducted with three primary school teachers in Saudi Arabia, who teach Science. The aim of this interview was to extract knowledge of lesson design and delivery of tutorials similar to the traditional classroom delivery model. The teachers evaluated the designed tutorial in order to give feedback and approval. Once the tutorials were approved by the teachers, the tutorials were implemented within LANA-I using a relational database MySQL to store and retrieve the resources. The LANA-I knowledge base consists of two layers: the domain, which is the science tutorials layer, and a general context layer, illustrated in Figure 4.

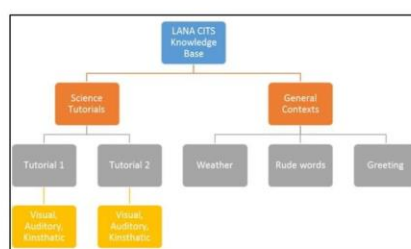


Fig. 4. Knowledge base layers

In figure 4, each layer represents a context, and each context has all related sub contexts mapped to it. The science tutorials hold all the tutorials sessions related to the science subject such as earth, moon and solar system. General contexts deal with general conversation not related to the domain, such as greeting, weather, rude words, and user leaving words (any words or sentences means that the user will leave the conversation). The general context makes LANA-I seem more aware by responding to the user's utterances, which are not related to the main domain. However, only a few general sub-contexts have been implemented as a proof of concept within the scope of this research.

After the tutorial was designed and approved by teachers then the tutorial questions were mapped to the VAK learning style model. Each question in the tutorial is mapped to Visual learning style through pictures and videos, Auditory learning style through sound recording, Kinaesthetic learning style through models and instructions as shown in Table 1.

Table 1. Part of the tutorial and how it is mapped to the VAK model

Tutorial	VAK
LANA: The solar system is the system, which comprising the sun and the objects that orbit it. Do you know what objects that exist in the solar system? User: Earth, moons and other planets	V: Picture A: Audio K: solar system model
IANA: Object moves in the gravitationally curved path called the orbit. The earth rotates around the sun and around its own axis. Do you know the time that the earth takes to rotate around its axis? User: A day, or 24 hours.	V: video A: Audio K: earth-moon model
LANA: What the result of this rotation? User: The day and night	V: Picture A: Audio K: earth-moon model

2.2 Arabic Scripting Language

The three different approaches to develop an Arabic CA and a number of challenges faced by Arabic language were discussed in survey paper [6]. It was concluded that there is a lack of Arabic NLP resources leading to limited capabilities within Arabic CAs. Consequently, this research needs a new scripting language to enable the Arabic CA combined with an ITS to deliver Arabic tutoring sessions using the modern standard Arabic language.

The foundation of LANA-I scripting language is based on the InfoChat scripting language [9]. InfoChat was designed using English scripting language and based on the pattern matching (PM) technique, where the domain is organised into a number of contexts and each context contains rules, each rule in the domain contains a number of patterns and a response that forms the CA output to the user. In the LANA-I CA engine, the similarity strength is calculated through combining the use of the PM, which is based on InfoChat method, and the Cosine algorithm [10] to compute the similarity strength between the utterance and the scripted patterns in order to improve the CA accuracy. The highest matching strength pattern will generate the response back to the user. Based on the scripting methodology reported by Latham [11] the procedures to create the scripts within the Knowledge base are:

1. The methodology for scripting each context is:
 - Create a context table, which has a record with a unique name to represent that context (topic).
 - When the context is invoked, an initialisation rule is fired.
 - All rules and patterns are scripted for the associated context.
 - Test each context to check that the expected rule is fired for a sample of user utterances.
2. The methodology for scripting each rule is:
 - Create a rule table, which has a unique rule name to represent the tutorial question.
 - For each rule, create patterns that match user utterances. Extract the important words and use the wildcards to replace unimportant words.

- Create patterns from each different utterance phrase. E.g. saying the same things using alternative words.
- Create the CA response to the utterance.
- Add Image, Audio, and instructions to the rule in order to map the learning style.

2.3 Scripting Arabic CA for LANA-I

In LANA-I, the contexts represent the tutorial topics and the rules represent the agent's questions for such topic, while the pattern represent the user's utterances, which belong to such a rule. The scripting language in LANA-I includes the following features:

- Provide supporting material to the user depending on the user's learning style (Visual: images and videos – Auditory: Sounds – Kinaesthetic: Instructions and objects). The learning style will be determined using a child friendly-customised VAK learning style preferences questionnaire before the tutorial.
- This material is stored in the scripting database and once a rule is fired, the corresponding material is provided to the user through the interface. All images, videos and audios provide the right answer. All of them has the same knowledge that the written text provides.
- The scripting language works with the ITS manager to check the user's learning style and provide the suitable material with the fired rule.

Table 2 shows an example of the scripting language and how the tutorial mapping the VAK model. When the user is visual learner, the rule is fired with the video or image material. In this example (Table 2), the rule has video that will be executed along with the rule. When the user is an Auditory learner, the rule will be fired with the audio material. The same thing with the kinaesthetic learning style, the rule will be fired with instructions to use the models. Figure 5 shows the models that are used with the Kinaesthetic learning style.

Table 2. Scripting language (translated)

Context: I Context Name: Topic 1 (Earth, Sun, and Moon)	Context: I Context Name: Topic 1 (الأرض، الشمس، والقمر)
Rule Number: 1 Rule Name: Object moves in the gravitationally curved path called the orbit. The earth rotates around the sun and around its own axis. How long the earth takes to rotate around its axis?	Rule Number: 1 Rule Name: يتكون النظام الشمسي من مجموعة أجرام مثل الأرض والقمر والكواكب. تتحرك الأجرام حول الشمس في مسار يسمى بالمدار. تدور الأرض حول الشمس وتدور حول محورها فهل تعرف كم تستغرق الأرض لتدور حول محورها؟
Pattern: It * a day Pattern: It * one day Pattern: It * 24 hours Pattern: * twenty four hours	Pattern: يوم واحد * Pattern: يوم كامل * Pattern: 24 ساعة * Pattern: * أربع وعشرون ساعة
Response: Excellent, the earth takes one day (24 hours) to rotate around its axis Image: yes Audio: yes Instruction: Look at the solar system model and its object	Response: أحسنت بارك الله فيك. تستغرق الأرض لتدور حول محورها 24 ساعة أي يوم كامل Image: yes Audio: yes Instruction: أنظر إلى نموذج النظام الشمسي والأجرام المحيطة بها

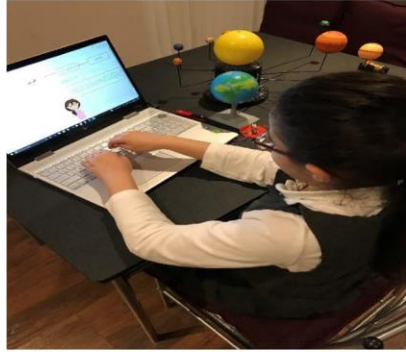


Fig. 5. The solar system model and earth model

2.4 Designing and Implementing the Arabic CA

The second component of LANA-I (Figure 3) is the Arabic CA. The Arabic CA architecture illustrated in Figure 1 was implemented, with the components developed using JAVA. The knowledge base and script databases were developed using the MySQL. The main components within the Arabic CA architecture (Figure 1) are: i) Controller, ii) Conversational Agent Manager, and iii) Conversational Agent Engine. The controller is responsible for directing and managing the entire conversation. The controller works with several other components, which are ITS manager, GUI and CA manager to achieve the conversation goal. First, the controller communicates with the ITS manager to find the student's learning style. Second, it gets the user's utterance and provides an utterance checking process. The controller will check the utterance based on the following constraints:

- Cleansing the utterance: The controller uses the utterance filter to remove special characters (i.e. \$, &, *, !, ?, "", £, (), ^) from the utterance. For example (as shown in English), if the user writes: (Hi, how are you Lana???) the system will convert it to (Hi how are you Lana).
- Check for rude and offensive words: the system will warn the user three times before terminating the session.

After the utterance is parsed the controller then works together with the CA to ensure that the conversation is within the tutorial scenario, or whether the user has switched context i.e. chooses to talk about something not related to the tutorial. Then, the controller delivers the response back to the user with the supporting material such as picture, video, audio, or instructions according to the user's learning style as shown in Figure 6.



Fig. 6. Conversation between LANA-I and Visual learner

The Conversation Manager controls the conversation to make sure that the goal is achieved, which is completing the tutorial. The conversation manager checks whether the user stays on the tutorial topic, or the user switches the context i.e. asks about something irrelevant to the tutorial. The Arabic CA implemented in LANA-I adopts a goal-oriented methodology[11]. For example, when the CA manager receives an utterance that is not related to the tutorial topic, the CA manager checks the user utterance with the general contexts layer to see if the utterance matches other general contexts within the database. If the utterance matches one of these contexts, the CA manager generates the response to the user and then the user is brought back to the point where the conversation was interrupted and then directed towards the goal.

The Conversational Agent Engine contains a combination of methods of string similarity and PM approaches to determine the similarity between two sets of strings within CA's, while traditional CA's used only a PM approach that involves a strength calculation through different aspects of the user utterance and the scripted pattern such as activation level and number of words etc. It is important to use string similarity methods to overcome some of the challenges in the Arabic language (described in section 2.3). In the field of CA development, the scripting is the most time consuming part of CA development [12]. The biggest challenge of scripting CAs is the coverage of all possible user utterances [11]. The engine handles the challenge of Arabic scripting by combining the Wildcard PM Function with the string similarity algorithm that calculates similarity strength and overcomes the scripting length challenge. Consequently, strings with minor differences should be recognized as being similar. The LANA-I CA engine has two main components that were specifically developed to deal with the features unique to the Arabic language in terms of its grammatical and morphological nature. These components are:

- Pattern matching approach (Wildcard PM [9]).
- String similarity algorithm (Cosine algorithm[10])

The combination of these components within LANA-I CA engine will help to calculate the similarity of the user utterance with scripted patterns using string similarity metrics. Consequently, reducing the need to cover all possible utterances when scripting the domain.

The first component of the Arabic CA engine, which is the primary phase of the engine, is the pattern matching. It is based on similar method to InfoChat [9] of PM, where the user utterance is matched to stored patterns. These patterns contain wild card characters to represent any number of words of characters. For example, the wild card symbol (*) matches any number of words, where the wild card symbol (?) matches a single character. An example for the PM approach in LANA-I is illustrated in the following table:

Table 3. Example of matching the utterance with the patterns script

Rule: Although the moon can appear a very bright white, its surface is actually dark. So from where the moon reflects the light?	Rule: القمر لا يضيئ بنفسه فهو انعكاس لضوء آخر. ماهو؟
Pattern 1: the moon reflects * sun Pattern 2: * the sunlight Pattern 3: * the sunshine	Pattern 1: القمر يعكس * الشمس Pattern 2: * ضوء الشمس Pattern 3: * نور الشمس
Utterance 1: it is the sunlight Utterance 2: reflects the moon the sun.	Utterance 1: هو ضوء الشمس Utterance 2: يعكس القمر الشمس

In Table 3, the utterance#1 will match pattern#2 (* the sunlight), where the wild card symbol (*) will match any number of words. The results from this matching will be 1. Whereas the utterance#2 will not match any patterns because of the word order. In this case the matching result will be 0.

The second component of the Arabic CA engine uses the STS algorithm, which for the purpose of this work is the Cosine similarity measure [10]. It determines the similarity between two pieces of text by representing each piece of text in the form of word vector. A word vector is a vector of length N where N is the number of different tokens in the text. The similarity is computed as the angle between the word-vectors for the two sentences in vector space.

For two texts t_1 and t_2 the cosine similarity is illustrated in the following equation:

$$SIM(t_1, t_2) = \frac{\sum_{i=1}^n t_{1i} t_{2i}}{\sqrt{\sum t_{1i}^2} \times \sqrt{\sum t_{2i}^2}} \quad (1)$$

The cosine similarity result is nonnegative and bounded between [0,1] where 1 indicates that the two texts are identical.

2.5 LANA-I CA Engine Workflow

First, the system will use the PM Wildcard to match the user utterance with the patterns saved in the database. If the match is not found, then the next step is to use the STS Cosine similarity. To illustrate the cosine similarity algorithm, assume the pattern stored in the LANA-I database was (S1), while the user utterance was (S2), as shown in Table 4:

Table 4. Example of calculating the similarity

Stored pattern in LANA-I: S1= يختلف القمر عن الأرض بأن ليس له غلاف جوي	S1: The moon is different from the earth that it does not have an atmosphere.
User Utterance: S2= القمر ليس له غلاف جوي لذلك يختلف عن الأرض	S2: The moon does not have atmosphere therefore it is different from the earth.

The utterance is not recognised by the PM approach because of the word order or minor differences from the pattern. Therefore, the system applies the Cosine similarity, which is illustrated in the following steps:

- Create Matrix[][] where the columns are the unique words from S1 and S2, and the rows are the words sequence of S1 and S2.
- Calculate the similarity between each word, 1 means the two words are identical, 0 means the two words are different.

The result obtained from Cosine similarity when applied to S1 and S2 is equal (0.88), which is greater than the threshold (0.80). This threshold is empirically determined using a small set of sentence pairs as in [13]. When the user's utterance is recognized by the similarity measure, the corresponding response will be generated and delivered to the user.

2.6 LANA-I ITS Implementation

This section describes the third component in the LANA-I architecture (Figure 3), which contains two sub-components (Figure 2), The Graphics User Interface (GUI), and The ITS manager.

The GUI is the point where LANA-I and the user interact with each other. LANA-I has a character called LANA (shown in figure 7), who appears in all the system interfaces in order to make it more engaging and natural for the users. Massaro and Bosseler [14] developed a computerized tutor called (Baladi) to teach vocabulary and functional language used to children with Autism with and without a cartoon face. They evaluated whether the face would increase the rate of learning for receptive measures. The results showed that the children's performance with the face condition was more significant than tutor with no face [14]. Based on that research, the LANA character was designed by the author and then evaluated by primary school teachers in Saudi Arabia in order to make the tutorial more engaging.



Fig. 7. LANA-I GUI and its character

The ITS manager is responsible for personalising the tutorial according to the user's learning style, which will be determined at the beginning of the tutorial through a questionnaire. The ITS manager can interact with the conversation agent components through the Controller, the Knowledge base, and the GUI.

In order to personalise the tutorial session according to the student's learning style, a questionnaire is applied first for each child. The questionnaire was developed on the basis of a widely disseminated version of [15] to be completed by pupil with parents or teacher help. This questionnaire focused on Smith's visual, auditory, and kinaesthetic styles (VAK). There were three questions for each style and the pupils had to respond 'yes' or 'no' to each question. The total question in the questionnaire are nine and the student's learning style result will be based on the highest score in one area.

2.7 LANA-I Workflow

This section describes the LANA-I workflow from perspectives of teacher and the pupil in order to understand how each activity communicates with others. Figure 8 presents the workflow.

Initially LANA-I starts from the learning style questionnaire, which is taken by the teacher. In the registration screen, the system asks the teacher to enter the pupil ID number (provided by the researcher), age, gender, and the result of the learning style questionnaire, conducted before using the system. There are four options for the learning style: Visual, Auditory, Kinaesthetic, and No-learning style, which used with the control group in the testing stage. After completing this stage, the system shows the pre-test interface, and asks the user to complete the test. The next interface after submitting the test is the CA tutorial. The ITS manager is responsible in this stage for personalising the tutoring session according to the user's learning style by providing the CA components, through the controller, the suitable materials from the Knowledge base component. The ITS manager also saves the user's registration information and the pre-test score in the log file/student's profile. In the CA screen, the user is led through the topic from one question to another until the end of the tutoring session. The user has only one attempt to answer. After completing the tutorial, the user is asked to fill the post-test questions and the score is saved by the ITS manager.



Fig. 8. LANA-I workflow

3 Experimental Methodology

The LANA-I prototype was tested through two main experiments to evaluate the system. The evaluation was based on a set of objective and subjective metrics [16]. The objective metrics were measured through the log file/temporal memory and the pre-test and post-test score. The subjective metrics were measured using the user feedback questionnaire. The aim of capturing the subjective and objective metrics is to test the hypotheses, which relate to the effectiveness of the conversation and using the VAK learning style, end user satisfaction, usability and system robustness. The main hypothesis of the experiments is:

- HA₀: LANA-I using VAK model cannot be adapted the student learning style.
- HA₁: LANA-I using VAK model can be adapted the student learning style.
- HB₀: LANA-I is not an effective Arabic CA.
- HB₁: LANA-I is an effective Arabic CA.

The aim of the first experiment was to test Hypothesis A (LANA-I using VAK model can adapt the tutorial to the student's learning style). The second experiment was conducted to test the Hypothesis B (LANA-I is an effective Arabic CA).

3.1 Participants

The total size of the sample was 24 neurotypical participants within the age group (10-12) years and their first language was Arabic. The participants recruited for the evaluation were residents of the Greater Manchester area within the UK and none of them had previous experience of using LANA-I. All participant's parents received an information sheet about the project and its aims, and a consent form to obtain their permission before conducting the experiment. The participants were divided into two groups. The first group is a control group (n=12), who used the LANA-I without adapting to the learning style VAK model as basis comparison. The second group is an experimental group (n=12), who used LANA-I with adapting to the learning style VAK

model. From 15 participants, 12 were selected based on their results on the learning style questionnaire in order to divide them into 3 groups of 4.

3.2 Experiment 1: LANA-I Tutoring with and without VAK learning Style

Subjective and objective metrics were used to answer the two questions related to Hypothesis A. The experiment is based on the pre-test and post-test scores with and without adapting to the VAK learning Style model. Participants were divided into 2 groups. Each group of participants was asked first to register into the system and complete the pre-test questions. The Control Group used LANA-I without adapting the VAK learning style model. They started the tutorial without the VAK questionnaire, whereas the experimental Group, who used the LANA-I with adapting to VAK learning style model, were asked to fill the VAK learning style questionnaire in order for LANA-I to be adapted to the learning style. After adapting the learning style, the tutorial provided the most suitable material during the session such as videos, images or instructions and physical resources. The Control Group did the tutorial session based on a text conversation without any additional materials. When the session ended, both groups did the post-test questions in order to measure their learning gain. Finally, both groups completed the user feedback questionnaire.

3.3 Experiment 2: LANA-I CA System Robustness

The data for this experiment was gathered from the LANA-I log file and the user feedback questionnaire whilst participants were completing experiment 1. The subjective and objective metrics were used to answer questions related to Hypothesis B. The data gathered from the log file allows assessment of the performance of LANA-I CA and the algorithms deployed in the architecture. This data will measure success using objective metrics. The data from user feedback questionnaire will be analysed in order to measure success using subjective metrics.

4 Results and Discussion

The learning gain was measured using a pre-test and post-test approach [17-19]. The same test questions were completed before and after the tutoring conversation. The test scores were compared to establish whether there is any improvement as follows:

$$\text{Learning gain} = \text{post-test score} - \text{pre-test score} \quad (2)$$

The results from experiment 1, illustrate that there is a statistically significant difference (as shown in Table 5) between the user's test scores with and without adapting to the VAK learning style. The mean rank of the differences between the pre-test and post scores in each case are without VAK ($M = 9.50$) and with VAK ($M = 15.50$). The P value of the mean's variation between the two groups is 0.03 (significant at $p < 0.05$), that meaning the users performed better when they are using their preferred VAK learning style.

Table 5. Mann-Whitney U test for the differences in Learning gain between the Control Group and the Experimental Group.

	Group	N	Mean Rank	Sum of Ranks
Learning gain	Control	12	9.50	114.00
	Experimental	12	15.50	186.00
	Total	24		
Test Statistics^a				
	Learning gain			
Mann-Whitney U				36.000
Wilcoxon W				114.000
Z				-2.173
Asymp. Sig. (2-tailed)				.030
Exact Sig. [2*(1-tailed Sig.)]				.039 ^b

The effect size (r) is calculated using the equation (3) and found that $r = 0.4$ indicating a medium to large effect size using [20] criteria of 0.1 = small effect, 0.3 = medium effect, 0.5 = large effect. It can be concluded that there is a statistically significant improvement in the learning gain in the Experimental Group comparing to the Control Group.

$$r = Z / \text{square root of } N \text{ where } N = \text{total number of cases (3)}$$

Relative learning gain was also measured in this experiment. Relative learning gain is a measure that calculate the average improvement in test scores as a percentage of the possible improvement [21]. This measure additionally takes into account the opportunity for improvement. Average test score improvements were calculated and compared using the following formula:

$$\text{Relative learning gain} = (\text{PostTest} - \text{PreTest}) / (\text{TotalScore} - \text{PreTest}) (4)$$

For example, if student get 8/10 in pre-test and only improves by 1, this is different to another getting 3/10 in pre-test and only improving by 1 – improvement is 50% in first case but only 14.2% in the second case. Table 8 illustrates the results of the Mann-Whitney U test conducted in order to measure the relative learning gain between Control Group and Experimental Group. It shows that there is a difference in the mean value between Control Group and Experimental Group. The average improvement in test scores in the Experimental Group were improved more than the Control Group. The ranks in each case are: Control Group ($M = 9.04$) and Experimental Group ($M = 15.96$). The relative learning gain between the Control Group and the Experimental Group is statistically significant different, p -value less than 0.05 ($P = .015$). The effect size (r) is calculated using the equation (3) and found that $r = 0.5$ indicating a large effect size using [20] criteria.

Table 6. Mann-Whitney U test for the relative learning gain between Control Group and Experimental Group.

Control Group and Experimental Group.					
		Group	N	Mean Rank	Sum of Ranks
Relative Learning gain		Control	12	9.04	108.50
		Experimental	12	15.96	191.50
		Total	24		
Test Statistics^a					
Relative Learning gain					
Mann-Whitney U					30.500
Wilcoxon W					108.500
Z					-2.428
Asymp. Sig. (2-tailed)					.015
Exact Sig. [2*(1-tailed Sig.)]					.014 ^b

This result highlight that the participants performed better when they are adapting to their preferred VAK learning style. It can therefore be concluded that there is statistically a significant difference between the pre-test and post-test scores with and without adapting to VAK learning style, indicating that the HA₁ can be accepted.

Another test was conducted to find out whether there was a significant difference in the scores for participant's perception of remembering what they have learned from LANA-I. The users who learned with adapting to VAK learning style are much happier with remembering what they have learned from LANA-I (91.7%) than users who learned without adapting to VAK (50%) as shown in Figure 9. This result indicates that HA₁ can be accepted.

Further tests have been carried out to find out whether there was a significant difference in the scores for participant's satisfaction with adapting the tutorial to the VAK learning style. The data was gathered from the user feedback questionnaire (Question 4: I think I can learn better with adapting to my learning style (Visual – Auditory -Kinaesthetic). The results illustrated that participants are happy with (83.3%), **indicating that HA₁ can be accepted.**

The results in Table 7 reveal that in general LANA-I performed well as an Arabic CA based on the number of corrected and uncorrected responses. The results show from the total conversations there are 281 correct responses (about 89.8 % from the total utterances) and 32 incorrect responses, when the algorithm failed to detect the students answer because of spelling mistakes (about 10.2% from the total utterances). The spelling mistakes affected the strength of similarity of the utterances, this cause the incorrect responses. Moreover, from the conversations that did reach the end of the tutorial, the majority (95.8%) of them reached the goal, meaning the user stayed within the tutorial and did not interrupt the conversation before the end of the tutorial. This provides evidence that the users did enjoy using LANA-I as a tutoring system, this is also reflected in the opinion of the participants in the end user questionnaire who thought that LANA-I is enjoyable (100%) and a vast majority stated they would use the system again (91.6%).

Table 7. Data analysis from the log file

CONVERSATION ANALYSIS	
METRIC	
Total number of conversations	24
Total number of utterances in all conversations	313
Average number of words per user utterance	3.1
Average number of utterances per conversation	31
Number of correct response from LANA	89.8%
Number of incorrect response from LANA	10.2%
Number of unrecognized utterances	38%
Number of utterances using Similarity Strength	57.18%
Number of rules fired	170
Average conversation duration (mins)	11.5
Percentage of conversations leading to goal	95.8%

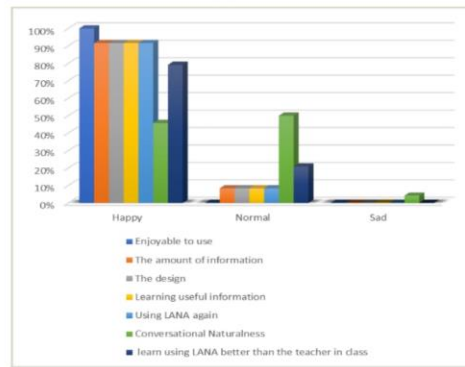


Fig 9. Frequency analysis LANA-I questionnaire

An additional insight provided from these results is that the LANA-I engine is allowing the reduction of scripted patterns. The results illustrated that (57.18%) of all the utterances input by the users were actually different from the scripted patterns and in this case the system used the Short Text Similarity algorithm (Cosine algorithm). The log file shows that 34 unique utterances relating to 15 different rules were correctly recognised and dealt with by the LANA-I algorithm by firing the appropriate rule. Based on these findings Hypothesis HB₁ can be accepted. However, the results have identified some weaknesses in the LANA-I architecture, mainly the number of unrecognised utterances. LANA-I failed to recognise some utterances from the users (38%). Upon further analysis of the log file it was found that some of these unrecognised utterances were due to the gaps in the knowledge base, which can be

easily addressed by expanding the patterns in the knowledge base. However, further research is needed to improve the similarity measure in order to reduce the number of patterns in the knowledge base.

5 Conclusion

This paper outlined a novel Arabic CITS called LANA-I, which used the VAK learning style model to enhance the learning of children. Our findings provide evidence for these novel features:

1. LANA-I can be adapted to the VAK learning style for the tutorial.
2. LANA-I is able to converse in the Arabic language to teach children age (10-12) the science subject.
3. The LANA-I engine used PM with STS similarity algorithm was effective in recognising the user utterance.

The pilot study illustrated some key information with regards to the effectiveness of using the VAK learning style with the tutorial and with regards the functionality and robustness of LANA-I CA. To summarise the main findings of the pilot study are as follows:

- Using VAK learning style with LANA-I improves the participants (neurotypical) learning.
- LANA-I can be adapted to the user's learning style and enhance their learning.
- LANA-I engine managed to reduce the number of scripted patterns by an average of 11% using the STS algorithm comparing to PM approach.
- LANA-I engine managed to respond correctly to the user by an average of 89%
- LANA-I is able to lead the user towards the goal of the tutorial, with 95.8% success of completing the tutorial.

However, the first evaluation highlighted areas of weakness within LANA-I architecture. Further research is required to make components and algorithms within LANA-I more robust and to achieve the main objective, which is developing an Arabic CITS for people with Autism. Additional research is required as follows:

- Further improvement to the knowledge base and the CA engine will be made based on the results of the first pilot study.
- New methodologies will be researched and developed to overcome the spelling variations in the Arabic language, which affect the performance of the similarity algorithm.
- Further enhancement to the similarity algorithm to improve the matching of Arabic text much more efficiently in order to reduce the number of incorrect responses.

These weakness and further enhancements will be addressed by further research and development, to make the system ready for use with autistic pupils.

References

1. TheNationalAutisticSociety. (2015). *Autism*. Available: <http://www.autism.org.uk/about-autism/introduction/what-is-autism.aspx>
2. A. A. Alotaibi, Knowledge and use of applied behavior analysis among teachers of students with Autism Spectrum Disorder in Saudi Arabia. Washington State University, 2015.
3. S. Fletcher-Watson, s. Hammond, A. O'Hare, A. M. Petrou, Helen, and H. Pain, "Click-East: Evaluating the Impact of an Ipad App On Social Communicative Abilities in Young Children with Autism," presented at the 2013 International Meeting for Autism Research, 2013.
4. V. Bernard-Opitz, N. Sriram, and S. Nakhoda-Sapuan, "Enhancing Social Problem Solving in Children with Autism and Normal Children Through Computer-Assisted Instruction," *Journal of Autism and Developmental Disorders*, vol. 31, no. 4, p. 377, 2001.
5. S. S. Aljameel, J. D. O'Shea, K. A. Crockett, A. Latham, and M. Kaleem, "Development of An Arabic Conversational Intelligent Tutoring System for Education of Children with ASD," in *Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), 2017 IEEE International Conference on*, 2017, pp. 24-29: IEEE.
6. S. S. Aljameel, J. D. O'Shea, K. A. Crockett, and A. Latham, "Survey of string similarity approaches and the challenging faced by the Arabic language," in *Computer Engineering & Systems (ICCES), 2016 11th International Conference on*, 2016, pp. 241-247: IEEE.
7. N. J. Ahuja and R. Sille, "A critical review of development of intelligent tutoring systems: Retrospect, present and prospect," *International Journal of Computer Science Issues*, vol. 10, no. 4, pp. 39-48, 2013.
8. M. o. Education. (2017). *Ministry of Education*. Available: <https://www.moe.gov.sa/en/Pages/default.aspx>.
9. D. Michie and C. Sammut, "Infochat scripiter's manual," *Convagent Ltd., Manchester*, 2001.
10. G. Qian, S. Sural, Y. Gu, and S. Pramanik, "Similarity between Euclidean and cosine angle distance for nearest neighbor queries," in *Proceedings of the 2004 ACM symposium on Applied computing*, 2004, pp. 1232-1237: ACM.
11. A. Latham, K. Crockett, D. McLean, and B. Edmonds, "A conversational intelligent tutoring system to automatically predict learning styles," *Computers & Education*, vol. 59, no. 1, pp. 95-109, 2012.
12. J. O'Shea, Z. Bandar, K. Crockett, and D. McLean, "A comparative study of two short text semantic similarity measures," in *KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications*, 2008, pp. 172-181: Springer.
13. Y. Li, D. McLean, Z. A. Bandar, J. D. O'Shea, and K. Crockett, "Sentence similarity based on semantic nets and corpus statistics," *IEEE transactions on knowledge and data engineering*, vol. 18, no. 8, pp. 1138-1150, 2006.
14. D. W. Massaro and A. Bosseler, "Read my lips: The importance of the face in a computer-animated tutor for vocabulary learning by children with autism," *Autism*, vol. 10, no. 5, pp. 495-510, 2006.
15. A. Smith, *Accelerated learning in practice*. A&C Black, 1998.

16. J. O'Shea, Z. Bandar, and K. Crockett, "Systems Engineering and Conversational Agents," 2011, vol. 10, pp. 201-232.
17. D. Kelly and B. Tangney, "Adapting to intelligence profile in an adaptive educational system," *Interacting with computers*, vol. 18, no. 3, pp. 385-409, 2006.
18. N. K. Person, "AutoTutor improves deep learning of computer literacy: Is it the dialog or the talking head?," *Artificial intelligence in education: Shaping the future of learning through intelligent technologies*, vol. 97, p. 47, 2003.
19. J.-L. Lee, G. Orwig, G. Gunter, and L. Witta, "The effect of cognitive styles on a visually-oriented task in online learning environments," in *Society for Information Technology & Teacher Education International Conference*, 2006, pp. 3489-3494: Association for the Advancement of Computing in Education (AACE).
20. J. Cohen, "Statistical power analysis for the behavioral sciences 2nd edn," ed: Erlbaum Associates, Hillsdale, 1988.
21. A. Latham, K. Crockett, and D. McLean, "An adaptation algorithm for an intelligent natural language tutoring system," *Computers & Education*, vol. 71, pp. 97-110, 2014.

Enhanced Learning Skills for Children with ASD in Mainstream Schools: Using an Adaptable VAK Conversational Intelligent Tutoring System

Sumayh S. Aljameel, James D. O'Shea, Keeley A. Crockett, Annabel Latham, Mohammad Kaleem, *School of Computing, Mathematics and Digital Technology, Manchester Metropolitan University*

Abstract. This article describes the design, development, and evaluation of LANA II CITS. LANA II is a novel Arabic Conversational Intelligent Tutoring System (CITS) that adapts to the autistic children learning style Visual, Auditory, Kinesthetic (VAK) model. Previously, the first prototype (LANA I) has been developed and evaluated with neurotypical children in order to evaluate the effectiveness and robustness of the system engine. In this paper, the improved prototype LANA II is developed to overcome the weaknesses identified through evaluation which include grammatical features and the morphological nature of the Arabic language. A morphologic pre-processing feature and an enhanced Cosine similarity were added in order to overcome the issues with Arabic language. In addition, TEACCH approach that are used to educate autistic children, was adapted in the GUI. This paper also presented the evaluation methodology and the experiment used to evaluate the enhanced component of LANA II architecture. Two experiments were conducted, the first experiment with neurotypical children in order to compare the results from LANA II with the results from LANA I. The results from this experiment illustrated that the enhancements made to the LANA II architecture have made a statistically significant impact on the effectiveness of LANA II engine when compared to LANA I. The second experiment was conducted with Autistic children studying in Arabic mainstream schools. This experiment was conducted to evaluate the autistic student's perceptions and the learning gain in order to answer the research question: Can LANA II using VAK model be adapted to the autistic student learning style?. The results from this experiment indicated a statistically significant improvement between autistic students learning gain with and without adapting to the learning style indicating that LANA II can be adapted to the autistic children learning style and enhance their learning.

Keywords. Autism Spectrum Disorder; Arabic language; Conversational Agent; Intelligent Tutoring System; Pattern Matching; String similarity; VAK learning style.

INTRODUCTION

Despite the large number of students with Autism Spectrum Disorder (ASD) attending mainstream schools, the process of facilitating their learning and participation remains a complex and poorly understood area of education (Waddington and Reed, 2017). The limited research base in this area indicates that school is a stressful and anxiety-provoking place for many such students with social isolation, loneliness and bullying commonplace (Roberts and Simpson, 2016). Furthermore, although many teachers in mainstream schools are firmly committed to the principles of inclusive education, they do not feel that they have the necessary training and support to provide adequately for students with ASD (Lindsay et al., 2014). Two recent reviews have identified teaching strategies and approaches for students with ASD in mainstream schools as a key 'gap' in the knowledge base for special educational needs provision (Meng, 2017; Barnhart, 2017). One solution is to use a technology-based learning as a virtual tutor to supplement the education of children with ASD in mainstream schools. Several studies reported the benefits of technology-based learning and people with ASD (Grynszpan et al., 2014; Fletcher-Watson, 2014). The virtual tutor offers a structured environment that is preferred by people with autism as they like routine and

repetitive behaviours (Boucenna et al., 2014). Virtual tutors benefit to ASD children, who have difficulties in the social interactions (Parsons and Cobb, 2011).

Intelligent Tutoring Systems (ITS) include any computer program used in learning and contains intelligent technologies to personalise learning according to individual student characteristics, such as knowledge of the subject, emotion, and learning style (Dermeval et al., 2017). ITS provide instructions and feedback to the learners by performing their main tasks based on Artificial Intelligence methods (Perikos et al., 2017). An extension of an ITS is Conversational Intelligent Tutoring Systems (CITS), which use natural language interfaces rather than menus to engage with a learner (Latham et al., 2014). CITSs allow users to explore and learn topics through conversation and to obtain knowledge in a similar manner as a classroom i.e. through student teacher interaction. CITSs have been developed for different domains and the user's level. BEETLE II (Dzikovska et al., 2014) is one example of CITS, which teaches conceptual knowledge about basic electronics and electricity through guided experimentation. AutoTutor (Graesser, 2016) is another example of CITS, which is a computer tutor designed for students within universities to learn an introductory computer course. However, the main limitation of both BEETLE II and AutoTutor is that they do not consider the student's learning styles during a tutoring session. Oscar CITS (Latham et al., 2014) overcomes AutoTutors limitation by predicting and adapting dynamically to a student's learning style. In the Arabic language, CITS is considered as a new area of research. Abdullah CITS (Alobaidi et al., 2013) is a recent development of an Arabic CITS for modern Islamic education, which aims to teach children aged 10 to 12 years old essential topics in Islam. Similar to AutoTutor, Abdullah CITS does not consider the student learning styles. To our knowledge, no academic research exists on the Arabic conversational intelligent tutoring system for Autistic children except that proposed in this research (S. S. Aljameel et al., 2017).

LANA I (S. Aljameel et al., 2018; S. S. Aljameel et al., 2017) is a novel Arabic CITS, which delivers topics related to the science subject by engaging with the user in Arabic language. The Visual, Auditory, and Kinaesthetic (VAK) learning style model is used in (LANA I) to personalise the lesson according to the user learning style preference. The aim of the research presented in this paper is to design and implement an Arabic CITS for children with autism aged 10-12 using the (VAK) model. Part of the challenge in building such a system is the requirement to deal with the grammatical features and the morphological nature of the Arabic language. The research into Conversational Agent (CA) development techniques revealed that hybrid approach was the most appropriate approach to develop an Arabic CA (S. S. Aljameel et al., 2016). The engine of (LANA I) is based on the two main CA development strategies, A Pattern Matching (PM) and a Short Text Similarity (STS) that calculate the matching strength of a pattern to the user utterance. The two parts of the engine work together in order to overcome some of the language unique challenges of the Arabic language and to extract the responses from resources in particular domain, which is Science topics. LANA I was evaluated by neurotypical students (S. Aljameel et al., 2018). This paper presents the development and the evaluation of LANA II. The main novel contributions of this paper are:

1. Presentation of a new architecture (Improved from LANA I) for creating an Arabic CITS.
2. A new LANA Conversation Agent (CA) algorithm that reduces the scripting effort and unrecognised utterances by processing the spelling mistakes and morphological nature of the Arabic language, which were issues in LANA I.
3. Development a new CITS called LANA II, which adapts the TEACCH method that is used to educate autistic children. LANA II is the first CITS that mapped TEACCH method to the user interface.
4. A new general evaluation framework has been developed and tested from the objective and subjective perspective.
5. Evaluation of LANA II through two empirical evaluation experiments which validate the architecture components and the VAK learning style model with both neurotypical and Autistic students in mainstream schools.

In this paper two research questions were explored:

- Can an Arabic Conversational Intelligent Tutoring System (CITS) be adapted to the learning styles for autistic children?
- Do the enhancements made to the (LANA I) architecture improve the overall effectiveness of LANA II engine?

LANA CITS WITH IMPROVED ARCHITECTURE

This section presents the development of LANA II that address the issues and the weaknesses based on the findings from the evaluation of the first prototype (LANA I) (S. S. Aljameel et al., 2017; S. Aljameel et al., 2018). There were a number of areas highlighted where the components of LANA I could be improved. The issues that were highlighted in the evaluation of LANA I were mainly due to grammatical features and the morphological nature of the Arabic language. Other issues that were identified were the need to further expand the knowledge base in order to increase the naturalness of the conversation. The weaknesses that were made apparent through the end user evaluation and the log file are as follows (S. Aljameel et al., 2018):

- Arabic Language Features: Common spelling mistakes.
- Architecture Features: Cosine algorithm similarity calculation.
- Gaps in the knowledge base: Unrecognised utterances.

As well as addressing the weaknesses in LANA I system, this section details the development of the TEACCH method (Mesibov et al., 2003) that is used to educate autistic children and how this method could be effectively implemented into LANA II. The research and development decisions made in order to mitigate these weaknesses are as follows:

- To overcome the issue of common spelling mistakes and variations, a morphologic pre-processing feature was developed for the LANA II architecture.
- Cosine algorithm similarity calculation: the similarity algorithm (used to match inputs with responses in LANA I) was improved in order to recognise and better deal with the words that have prefix, suffix or differences in the word composition.
- Knowledge base expansion: The knowledge base was expanded in order for LANA II to seem more natural during conversation. In addition, more domain specific (i.e. Science topics) as well as general knowledge were added to the database.
- User Interface using TEACCH method: to make LANA II suitable for autistic children, TEACCH method was applied in the user interface.

These improved components are detailed in the following sections in this paper. The combination of these changes contributes to improving the effectiveness and accuracy in LANA II. Figure 1 illustrates the new architecture of LANA II. The figure outlines all components presented in this paper and how they interact with each other.

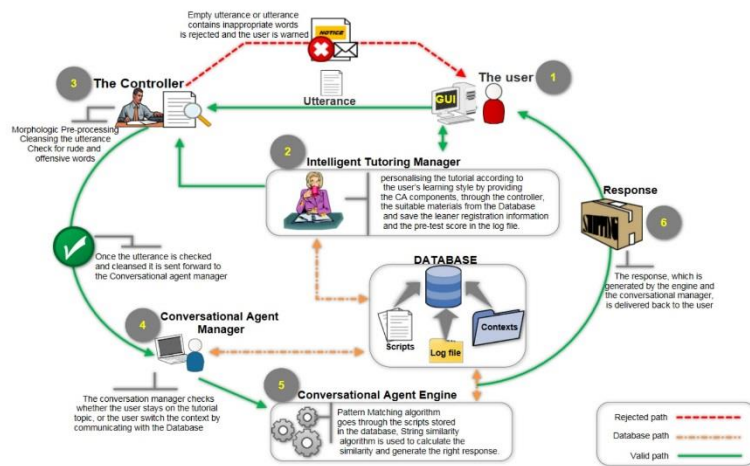


Fig 1. The new architecture of LANA II

Morphologic Pre-processing

One of the issues highlighted in the evaluation of LANA I that needed to be addressed is the common spelling variations and mistakes made by the user. These mistakes are due to the phonological similarity of some of Arabic alphabet characters. In the Arabic language, the letters change form according to their position in the word (beginning, middle, end and separate). Some of these letters undergo a light modification in writing which does not influence considerably the meaning of the word. For example, the letter "ا" at the beginning of a word can take different forms: "أ", "إ", "و" or "أ". A typical example is found in Arabic word (أربعة) instead of (أربعة), which is the correct spelling of the word (four). In this example, the two words have the same meaning but the first word has a common spelling mistake where each of the letter "ا" and "و" has a different form in the two words. In order to overcome the problem of the representation variation of Arabic letters, pre-processing methods were added to the controller and applied on the utterance before applying the similarity algorithm:

- Replacing "أ", "إ", "و" by Alif bar "أ"
- Replacing "أ", "و", "ي" by "أ" at the end of the words.
- Replacing "أ", "و" by "أ" at the end of the words.

The pre-processing method is applied by the controller (see figure 1). The controller is responsible for:

- Filtering the utterance from the special characters (i.e. \$, &, *, !, ?, "", £, (), ^)
- Check for rude words.
- Replacing the characters that caused spelling mistakes.

The expected impact of morphologic pre-processing will be to reduce the unrecognised utterances due to the common spelling variations and mistakes made by the user. In addition, this feature will reduce the number of scripted patterns since it will be no need to cover all letters variation in one word.

Cosine similarity

The Cosine similarity (Qian et al., 2004) was used in LANA I (S. S. Aljameel et al., 2017; S. Aljameel et al., 2018) in order to calculate the similarity between two texts. It solves the complex word order issue that comes with the Arabic language. The similarity is determined between two pieces of text by representing each piece of text in the form of word vector. A word vector is a vector of length N where N is the number of different tokens (i.e. words) in the text. The similarity is computed as the angle between the word-vectors for the two sentences in vector space. This means that the cosine compares the two texts using the word level approach. For example, the texts (الفصول) and (أربعة فصول) have the same meaning but they are different because of the prefix (أ) and the suffix (ة). In this example, if the cosine algorithm is applied, the similarity results will be zero. Therefore, in order to enhance the similarity calculation, the cosine similarity was adjusted to use the character level approach. This means that each word will be compared with other word character by character after applying the Bi-gram method (Kondrak, 2005). The general process of applying the enhanced cosine algorithm is illustrated in the following steps, where S₁ is the first utterance and S₂ is the second utterance:

S₁: الفصول الأربعة (four seasons with the prefix (أ) and the suffix (ة))

S₂: أربعة فصول (four seasons without the prefix and the suffix)

1. Create Matrix[[]] where the columns are the unique words from S₁ and S₂, and the rows are the words sequence of S₁ and S₂.
2. Calculate the similarity between each word by applying the bi-gram method first. It is difficult to compare the letters of the two words because some words have the same letters but different meaning. For example, the two words (rose, sore) have identical letters but they are different in the meaning. In the bigram, each word is represented by a list of its constituent n-grams, where n is the number of adjacent characters in the substrings, n=2 in the bigram. Using these lists, similarity measures between pair of words are

calculated based on cosine similarity approach. Table 1 shows an example of calculating the similarity of two utterances.

Table 1 Calculate the similarity between pair of words

Matrix [I][J]=					
فصول (Seasons) without prefix	اربع (Four) without prefix and suffix	الاربعه (Four) with the prefix (ال) and the suffix (هـ)	الفصول (Seasons) with the prefix (ال)		
0.77	0	0.2	1	الفصول (seasons) with the prefix (ال)	S ₁
0	0.71	1	0.2	الاربعه (four) with the prefix (ال) and the suffix (هـ)	
0	1	0.71	0	اربع (four) without prefix and suffix	S ₂
1	0	0	0.77	فصول (seasons) without prefix and suffix	

For example, the words 'فصول' and 'الفصول' have the same meaning (Four Seasons) but they are different because the second word has the prefix 'ال'. If the cosine similarity is applied directly the similarity between these words will be zero. In the new algorithm, the bigram is applied first between the words, in this example the bigram will be:

Word 1: فصل, صور, ول
Word 2: ال, لف, فصل, صور, ول

Then the cosine similarity is applied between these words:

Union: ال, لف, فصل, صور, ول

Vector1: 1,1,1,0,0
Vector2: 1,1,1,1,1

The cosine similarity is calculated using equation (1), where t_1 is vector1 and t_2 is vector2. The similarity between the two words is 0.77. Table 1 shows the similarity measures between each two words after applying the bigram method and the cosine similarity.

$$SIM(t_1, t_2) = \frac{\sum_{i=1}^n t_{1i} t_{2i}}{\sqrt{\sum t_{1i}^2} \times \sqrt{\sum t_{2i}^2}} \quad (\text{Eq 1})$$

- The last step here is to write the vector for each string when the similarity between the words is greater than 0.40, this threshold is empirically determined using a small set of sentence pairs.

$t_1 =$ 1 1 0.71 0.77

$t_2 =$ 0.77 0.71 1 1

Then the Cosine similarity measure is applied equation (1): $SIM(t_1, t_2) = 0.96$. Table 2 shows a pseudo-code representation of the new string similarity algorithm.

Table 2 A pseudo-code representation of the LANA II string similarity algorithm

Union= Unique words (Pattern AND Utterance)
String1= Words sequence of Pattern
String2= Words sequence of Utterance
Calculate the String1 vector
FOR EACH (Union) AND (String1)
{
IF (Word Similarity (String1, Union) is TRUE)
String1 vectors =1
ELSE
String1 vectors =0
}
Calculate the String2 vector
FOR EACH (Union) AND (String2)
{
IF (Similarity (String2, Union) is TRUE)
String2 vectors =1
ELSE
String2 vectors =0
}
COSINE SIMILARITY (String1 vectors, String2 vectors)
IF COSINE SIMILARITY >=80 THEN
Similarity=TRUE
ELSE
Similarity=FALSE

The new similarity algorithm explained above, significantly reduces the number of scripts that have to be scripted to deal with the issue of word structure (adding the prefix or suffix) an example of this is illustrated in Table 3. In Table 3, the scripted pattern (الفصول الأربعة) has five possible patterns. When applying the new algorithm, which is explained in Table 2, the scripted pattern will cover all five patterns, which may be different from the scripted pattern because of the prefix, suffix and word order.

Table 4 shows the differences in the similarity scores between the user's utterance and the scripted pattern using LANA I and LANA II algorithm. LANA I algorithm uses the pattern matching and Cosine similarity (the original), where LANA II algorithm uses Morphologic Pre-processing feature and the improved Cosine similarity. As shown in the table, it is noticed that LANA II performed better when the two strings are different because of the word structure such as the prefix or spelling mistakes such as duplicate the letter. In addition, LANA II performed better when there are differences in the word composition such as the words اختلاف and يختلف (difference, differ) noun and verb.

Table 3 The number of scripts patterns for the pattern (Four seasons)

Scripted pattern	الأربعه (four) with the prefix (ال) and the suffix (هـ)	الفصول (seasons) with the prefix (ال)
Patterns covered	فصول (seasons) without prefix	اربع (four) without prefix and suffix
	اربع (four) without prefix and suffix	فصول (seasons) without prefix
	الأربعه (four) with the prefix (ال) and the suffix (هـ)	فصول (seasons) without prefix
	فصول (seasons) without prefix	الأربع (four) with the prefix (ال)
	الأربع (four) with the prefix (ال)	الفصول (seasons) with the prefix (ال)

Table 4 Comparison of the similarity algorithm scores between LANA II and LANA I

Reason	Cosine_ word level	Cosine_ Character level	Utterance	Pattern	Question
Prefix and suffix	0	0.98	الأربع فصول	الفصول الأربعة *	تدور الأرض حول الشمس كل سنة. ماهي نتيجة هذا الدوران؟
Word composition and prefix	0	0.90	اختلاف الفصول السنة	تختلف فصول السنة	
prefix	0.23	0.89	ليس للقمر غلاف جوي	القمر ليس له الغلاف الجوي	لقمر هو أقرب الاجرام الفضاء الى الأرض ولكنه يختلف عن الأرض. ما هو وجه الاختلاف عن الأرض
Word composition and prefix	0.50	0.91	يختلف بعدم وجود هواء	الاختلاف بعدم وجود الهواء	
prefix	0.55	0.82	القمر لا يوجد به الأكسجين	لا يوجد به غاز الأكسجين	
prefix	0	0.92	انهار واليل	الليل والنهار *	تستغرق الأرض لتدور حول محورها 24 ساعة أي يوم كامل ماذا ينتج عن هذا الدوران؟
Duplicate the letter	0.52	1	الليل والنهار	الليل والنهار *	
prefix	0	0.98	كواكب	الكواكب *	

Knowledge base expansion

The knowledge base was expanded to address the finding from the end user questionnaire and the log file from the evaluation of LANA I (S. Aljameel et al., 2018). The participants expressed that the naturalness in LANA I conversation is low. In addition, the results from the evaluation of LANA I found that there are 38% unrecognised utterances because of the knowledge base gap. This means there are some utterances that LANA I could not recognise because of the gap in the knowledge base. Therefore, the knowledge base was expanded to make LANAI more natural in term of conversation and has a wider range of knowledge in general topics.

The domain specific knowledge (Science subjects) was increased by adding new contexts. This process involved interviews with three primary school teachers in Saudi Arabia, who teach the science subject. The knowledge base was expanded to include another tutorial related to the science topics. In addition, the first tutorial was expanded and all the unrecognised utterances resulting from weakness in the knowledge base from the evaluation of LANA I were added as new patterns to the knowledge base. Furthermore, the general topics of the knowledge base was expanded to include more topics and patterns such as greeting, weather and sport.

Mapping the tutorial into VAK learning style model

The second tutorial was designed in accordance with the methodology described in (S. S. Aljameel et al., 2017; S. Aljameel et al., 2018). After the tutorial was designed and approved by the teachers then the tutorial questions were mapped to the VAK learning style model. In order to personalise the tutorial session according to the student's learning style, a questionnaire will be applied first for each child. The questionnaire was developed on the basis of a widely disseminated version of (Smith, 1998) to be completed by student with parents or teacher help. This questionnaire focused on visual, auditory, and kinaesthetic styles (VAK). There were three questions for each style and the students had to respond 'yes' or 'no' to each question. There is a total of nine questions and the student's learning style result will be based on the highest score in one style.

Each tutorial question was associated to the VAK learning style material. Table 5 shows how the tutorial questions for the second topic is mapped to the learning style model VAK. Each question in the tutorial was mapped to Visual learning style through pictures and videos, Auditory learning style through audios, Kinaesthetic learning style was modelled through models and instructions.

Table 5 The second Tutorial: Electrical circuit and the magnetism

Question 1: Electricity has a big impact in our lives; we use it to the light of the lamp. Electricity is generated from electric charges, which are very small and cannot be seen. Electric charges are divided into two types: negative charges and positive charges. When a charged object comes near to another object, they will either attract or repel each other. What happens when similar charges converge (positive or negative)? Answer: If the charges are opposite - they attract	V: video A: Audio K: Electrical circuit.
Question 2: If we rub the balloon on the woolly object and then hold the balloon up on the wall. What will happen? Answer: The balloon stay stuck on the wall.	V: Picture A: Audio K: Electrical circuit.
Question 3: Electrical charges can be passed through some materials to be power supply. The power supply only works in a closed path. What is the name of this path? Answer: Electrical circuit.	V: Picture A: Audio K: Electrical circuit.
Question 4: The circuit consists of three basic parts. what is it? Answer: energy source, an electrical load (device), a conductor (wire).	V: Picture A: Audio K: Electrical circuit.
Question 5: The power is applied when the circuit is closed which means it is connected. How can we control the flow of electricity in the circuit to switch on or off the lamp? Answer: By connecting an electric switch.	V: video A: Audio K: Electrical circuit.
Question 6: Magnet is a material used to attract certain objects. Do you know what objects are attracted by the magnet? Answer: Objects made of iron, nickel or cobalt.	V: Picture A: Audio K: Magnet.
Question 7: Magnet can pull or push some objects without touching it, because there is an area around the magnet showing the effect of its power. What is this area called? Answer : This area is called the magnetic field.	V: Picture A: Audio K: Magnet.
Question 8: The magnet consists of two poles. What are they? Answer: The north pole and the south pole.	V: video A: Audio K: Magnet.
Question 9: What happens when we if you bring two north poles together, or two south poles together in two magnets? Answer: They repel and the magnets push each other away.	V: video A: Audio K: Magnet.
Question 10: What happens when we bring a north pole and a south pole together in two magnets? Answer: They attract and the magnets may stick together.	V: video A: Audio K: Magnet.

Mapping TEACCH method to LANA II

TEACCH (the Treatment and Education of Autistic and related Communication Handicapped Children) is an education programs approach for people with ASD. This approach is developed by Eric Schopler at the University of North Carolina in 1972 (D'Elia et al., 2014). TEACCH is based on Structured Teaching technique for autistic students, which is used in UK schools since 1990 then the National Autistic Society start collaboration with the Division TEACCH (Mesibov et al., 2003). The aim of structured teaching is to enhance the teaching methods and the environment by considering the cognitive ability, needs and interest of the autistic students. In addition, the structured teaching element aims to organize the classroom and to make the teaching methods and style friendly for autism (Mesibov and Shea, 2010). TEACCH has four main components: physical structure, daily schedules, work system, and visual cues and instruction. Many advantages could be gained from using the Structured Teaching, as a method of delivering the curriculum. Such as, enhancing and facilitating the teaching and learning process. In addition, it can improve access to the curriculum for many students with ASD (Mesibov et al., 2003). The TEACCH main components was mapped to LANA II as follow:

- The physical structure: It involves arranging the materials, furniture and general surroundings of the students in order to enhance the learning environment. In LANA II, the physical structure is applied by enhancing the overall Graphics User Interface (GUI) such as improving the colours quality, and adding instructions.
- The work system: It aims to organise the activity by explaining the activity, telling the student the required time to complete the activity, and what will they gain after completing the activity. In LANA II, the character LANA will provide the student an explanation of each interface. For example, LANA in Figure 2 (a) explained the pre-test and why and how to do it.
- The daily schedule: It provides all activities and their order, which will conduct during the day. This helped the autistic students to understand clearly, what they have to do and give a predictable environment as they need. In LANA II, an activity timeline in each interface is added that shows the current activity (Figure 2 (b) and (c)).
- The visual structure: It provides the student with ASD with information on how to complete the task because the visual skills is usually strong in autistic students. Visual information can be very helpful in developing understanding for student with ASD. Visual information has three kinds, which were used in LANA II. The first type is the visual clarity, which engages the student with ASD by drawing their attention to important information. In LANA II the conversation is coloured whereby the black text indicates the question, the red text is the user utterance, and the blue text is the instruction from CITS (see Figure 2 (d)). The second type is the visual organisation, which is the way the tasks are organised to allow the student with ASD focus on the task and complete it. In LANA II, an activity timeline is used to help to organise the tasks within the tutorial. The third type is the visual instruction, which are the written cues that provide the students with ASD information of what is the next task and how to complete it. In LANA II, the instruction interface is used before each activity in order to give the student information about the next task (see Figure 2 (a)).

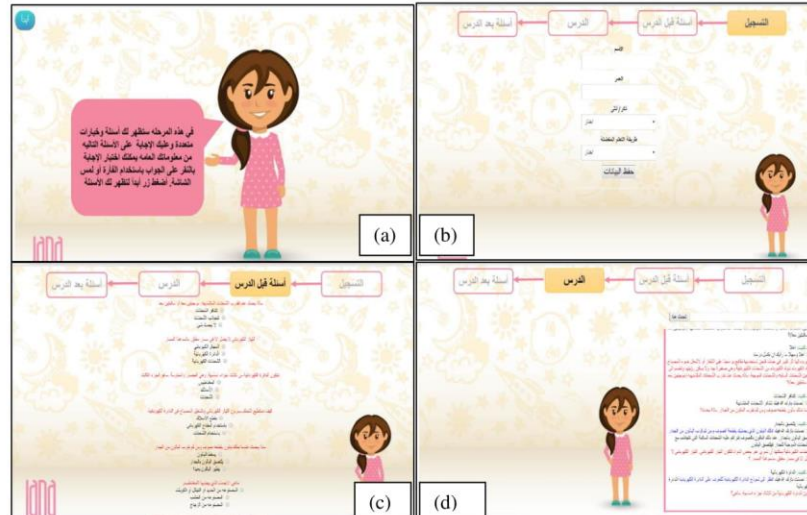


Fig 2. LANA II GUI (a) Applying the work system approach of TEACCH method, (b) Applying the scheduling approach of TEACCH method, (c) The activity timeline in the pre-test interface, (d) Visual instruction approach of the TEACCH method.

LANA II WORKFLOW

Figure 3 describes the system workflow. LANA II will start from the registration screen to enter the user name, age, gender, and the result of the learning style questionnaire, which was conducted before using the system. Then the system will show the pre-test interface, and ask the student to complete this test. After completing this stage, the CA tutorial will begin and the suitable materials will be provided to the student according to the student's learning style. If the student is from the control group, who will do the tutorial without learning style, the tutorial will be text-based only without any materials such as videos or images. In the CA screen the user will be led through the topic from one question to another until the end of the tutoring session. After completing the tutorial, the learner will be asked to fill the post-test questions, which are the same questions as the pre-test. Finally, feedback questionnaire will be provided to the student. The feedback questionnaire contains nine questions to be rated using the three faces scale (happy, natural, sad) (Massaro and Bosseler, 2006). It is reported by (Reynolds-Keefer and Johnson, 2011) that questionnaires with three face scales are more understandable with younger participants. Many children cannot give a verbal answer of their opinion or feeling, either because they are young, or because of a neurological disorder or communication challenging such as Autistic children (Yahaya and Salam, 2016). The questions were designed to extract the user's opinion on the performance and usability of LANA II by selecting the most appropriate face representing their feeling against each question.

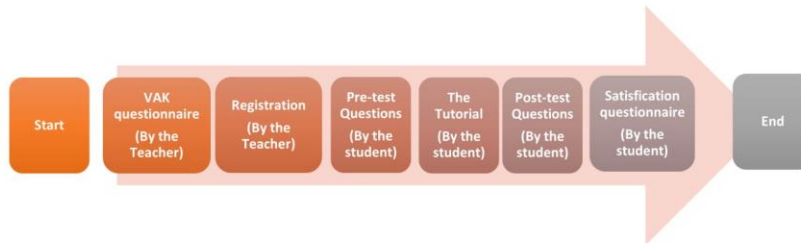


Fig 3. LANA II workflow

LANA II EVALUATION METHODOLOGY AND RESULTS

This section presents the design of an evaluation methodology in order to evaluate LANA II. Two experiments were conducted to evaluate LANA II. The first experiment was conducted with participants within the target age group (10-12) years old whose first language is Arabic. None of the children have been diagnosed as Autistic. This experiment is test the enhancement made in the LANA II engine to compare the metrics from the LANA I (S. Aljameel et al., 2018) and the metrics from LANA II. This will bring to light whether the enhancement made in LANA II components improve LANAs conversational ability. The second experiment was conducted with Autistic children on the target age group (10-12) years old whose first language is Arabic and have a high functioning autism (HFA) who have no language problem and intellectual impairments. This experiment was conducted to measure the capabilities from different aspects such as tutoring success, VAK learning style model and user evaluation. The main hypotheses of the experiments:

H0_A: the enhancement made to LANA II architecture did not improve the overall effectiveness of LANA II engine.

H1_A: The enhancement made to the LANA II architecture improve the overall effectiveness of LANA II engine.

H0_B: LANA II using the VAK model cannot be adapted to the Autistics learning style.

H1_B: LANA II using the VAK model can be adapted to the Autistics learning style.

Experiment 1: LANA II effectiveness

Participants

The total size of the sample was 24 participants with age group (10-12) years old and their first language was Arabic. The participants recruited for the evaluation were from UK and none of them had a previous experience using LANA II. All of the participant's parents received an information sheet about the project and its aims, and a consent form to get their permission before conducting the experiment. The participants were divided into two equal groups. The first group (12 participants) used the LANA II without the VAK learning style model. The second group (12 participants) used LANA II with the VAK learning style model.

Experiment Methodology

This experiment was conducted to test the hypothesis H1_A based on the metrics outlined in Table 6. The data for this experiment was gathered from the LANA II log file and the user feedback questionnaire. The subjective and objective metrics were used to answer questions related to Hypothesis H1 (The enhancement made to (LANA I) architecture improve the overall effectiveness of LANA II engine):

Question 1: Is the enhanced LANA II engine more effective than (LANA I) engine?

Question 2: Do the improvements made to the LANA II engine result in improved perceptions of LANA II by the users?

The tutorial that was given by LANA II is based on Science tutorials for children age group (10-12) years old. The experiment was undertaken in a classroom environment, and each student individually worked on a computer. The computer used in this experiment is touchscreen computer in order to let the students use both of the keyboard and the screen to interact with LANA II. The teacher was present during the experiment and the students were given instructions about using LANA II. During the tutorial session, all the questions from the system, the student answers, and the conversation variables/metrics were captured and saved in the log file, for further analysis to test the success of the tutorial session.

The data gathered from the log file will allow assessment of the performance of LANA II CA and the algorithms deployed in the architecture. This data will be compared with the set of data from the LANA I evaluation (S. Aljameel et al., 2018). This comparison will show if there is any statistically significant improvement between the two datasets measures.

Table 6 Experiment 1 metrics

Metrics to be Evaluated	Method of Evaluation	Metrics Type
LANA Tutor is enjoyable to use	Questionnaire	Subjective
LANA has a right amount of information in one Tutorial.	Questionnaire	Subjective
I like the design and colour of the interface.	Questionnaire	Subjective
I would like to use LANA again with different topics.	Questionnaire	Subjective
The conversation was natural.	Questionnaire	Subjective
I think I would like to learn using LANA better than the teacher in class	Questionnaire	Subjective
Percentage of correct response	Log file	Objective
Percentage of incorrect response	Log file	Objective
Percentage of unrecognized utterance because of the knowledge gap	Log file	Objective
Percentage of utterances using Similarity Strength	Log file	Objective
Percentage of utterances with correct response using STS algorithm	Log file	Objective
Number of the rule fired	Log file	Objective
Average conversation duration (minutes)	Log file	Objective

Experiment 1 Results

Log file analysis

The data from log file of LANA II was collated, processed and analysed. Table 7 summarises the results of data gathered through the log file during the evaluation of LANA I and LANA II. The results show that LANA II performed better during the end user evaluation. Further analysis was done for these results in the following section to determine whether these results are significantly different from the results of the first prototype LANA I.

Table 7 Log file analysis from comparison between LANA I and LANA II

CONVERSATION ANALYSIS		
All METRIC	LANA II	LANA I
Total number of conversations	24	24
Total number of utterances in all conversations	310	313
Average number of words per user utterance	3.1	3.1
Average number of utterances per conversation	13	31
Percentage of correct response	96.4%	89.8%

Percentage of incorrect response	3.6%	10.2%
Percentage of unrecognized utterance because of the knowledge gap	5.4%	38%
Percentage of utterances using Similarity Strength	42.7%	57.18%
Percentage of utterances with correct response using STS algorithm	42.7%	19%
Number of the rule fired	280	170
Average conversation duration (minutes)	12	11.5

Objective data analysis from LANA I and LANA II data sets

This section illustrates the objective data analysis to find out the differences in the data in order to test hypothesis H1_A (The enhancement made to the (LANA I) architecture improve the overall effectiveness of LANA II engine). Mann-Whitney U Test is used to test the significance of the results. Table 8 and Table 9 shows the results of the Mann-Whitney U test to test the objective data between LANA I and LANA II.

Table 8 Mean rank values from Mann-Whitney U test for the objective data between LANA I and LANA II

	Prototype	N	Mean Rank	Sum of Ranks
Number of unrecognised utterances	LANA I	24	36.50	876.00
	LANA II	24	12.50	300.00
Number of correct response	LANA I	24	17.67	424.00
	LANA II	24	31.33	752.00
Number of incorrect response	LANA I	24	30.56	733.50
	LANA II	24	18.44	442.50
Number of utterances with correct response using STS algorithm	LANA I	24	15.25	366.00
	LANA II	24	33.75	810.00
Number of rule fired	LANA I	24	12.50	300.00
	LANA II	24	36.50	876.00
Number of utterances using Similarity Strength	LANA I	24	27.08	650.00
	LANA II	24	21.92	526.00
Average conversation duration (mins)	LANA I	24	23.00	552.00
	LANA II	24	26.00	624.00

Table 9 P- Values from the Mann-Whitney U test for objective data between LANA I and LANA II

	Test Statistics ^a						
	Number of unrecognised utterances	Number of correct response	Number of incorrect response	Number of utterances with correct response using STS algorithm	Number of rule fired	Number of utterances using Similarity Strength	Average conversation duration (mins)
Mann-Whitney U	.000	124.000	142.500	.000	.000	226.000	252.000
Wilcoxon W	300.000	424.000	442.500	300.000	300.000	526.000	552.000
Z	-6.371	-3.911	-3.237	-6.395	-6.395	-1.369	-.864
Asymp. Sig. (2-tailed)	.000	0.000092	.001	.000	.000	.171	.388
a. Grouping Variable: Prototype							

Table 8 and Table 9 show that there is a statistically significant difference in the number of unrecognised utterances between LANA I and LANA II, p-values less than 0.05 (reported by the SPSS as p = .000). When comparing the

mean ranks, of the two systems LANA I was ranked higher, which means LANA I having more unrecognised utterances. The effect size (r) was calculated and found that $r = 0.9$ indicating a very large effect size using Cohen (1988) criteria of 0.1 = small effect, 0.3 = medium effect, 0.5 = large effect. Moreover, the number of correct responses between LANA I and LANA II was statistically significant ($p = 0.000092$). In comparing the mean rank LANA II ranks higher, which indicates that LANA II having more correct responses than the LANA I. The effect size (r) = 0.5 indicating a large effect size using Cohen (1988) criteria of 0.1 = small effect, 0.3 = medium effect, 0.5 = large effect. The results also indicated that there were statistically significant differences between the numbers of the incorrect response between LANA I and LANA II ($p = 0.001$). When comparing the mean ranks it can be seen that (ANA I ranks higher than LANA II, meaning that LANA I having more incorrect responses than LANA II. The effect size (r) = 0.4 indicating a nearly large effect size using Cohen (1988) criteria of 0.1 = small effect, 0.3 = medium effect, 0.5 = large effect. Furthermore, Table 8 and Table 9 show that there was a statistically significant difference in the correct responses using STS algorithm between LANA I and LANA II (reported by the SPSS as $p = .000$). When comparing the mean ranks it can be seen that the LANA II ranks higher than LANA I, meaning that the correct responses using STS algorithm in LANA II was significantly improved compared to the correct responses using STS algorithm of LANA I. The effect size (r) = 0.7 indicating a very large effect size using Cohen (1988) criteria of 0.1 = small effect, 0.3 = medium effect, 0.5 = large effect. In addition, the results indicated that there were statistically significant differences between the numbers of the rules fired between LANA I and LANA II (reported by the SPSS as $p = .000$). When comparing the mean ranks it can be seen that LANA II ranks higher than LANA I, meaning that LANA II having more fired rules than LANA I. The effect size (r) = 0.9 indicating a very large effect size using Cohen (1988) criteria of 0.1 = small effect, 0.3 = medium effect, 0.5 = large effect. However, when comparing the numbers of utterances using similarity strength between LANA I and LANA II, the results illustrate that there is not a statistically significant ($p = .171$) difference between LANA I and LANA II. Moreover, the results of the Mann-Whitney U test conducted in order to measure whether there was a statistically significant difference between the times taken to complete the tutorial between LANA I and LANA II. It can be concluded from the results that the time taken to reach the conversation goal between LANA I and LANA II was not statistically significant ($p = 0.388$). This concludes the results analysis of the log file data. The following section explores and analyses the questionnaire data that was gathered in order to gauge participants perceptions related to the subjective metrics.

Analysis of questionnaire data

Table 10 summarises the findings of the questionnaire from the evaluation of LANA I and LANA II. Table 11 shows the means values of the questionnaire one and two.

Table 10 Frequency analysis questionnaire data results from LANA I and LANA II

QUESTIONNAIRE RESULTS	Happy		Normal		Sad	
	LANA I	LANA II	LANA I	LANA II	LANA I	LANA II
I think LANA Tutor is enjoyable to use	100%	100%	0	0	0	0
I think LANA had a right amount of information in one Tutorial.	91.6%	91.6%	8.4%	8.4%	0	0
I like the design and colour of LANA.	91.6%	100%	8.4%	0	0	0
I think I remember what LANA has taught me in the tutorial.	91.7%	91.7%	8.3%	8.3%	0	0
I think I can learn better with using my learning style (Visual – Auditory - Kinaesthetic).	83.3%	91.6%	16.7%	8.4%	0	0
I think I would like to use LANA again with other lessons.	91.6%	91.6%	8.4%	8.4%	0	0
I think talking with LANA is like as talking with my friend.	45.8%	91.6%	50%	8.4%	4.2%	0
I think I would like to learn using LANA better than the teacher in class	79.16%	91.6%	20.84%	8.4%	0	0

Table 11 Mean values from evaluation questionnaire LANA I and LANA II

Mean							
Enjoyable	Satisfaction with the information.	Remembering and understanding	Satisfaction with learning style	Design and colour	Learning from LANA is better than the teacher.	Using LANA again with other lessons.	Naturalness
3.0000	2.9583	2.9583	2.9167	2.9583	2.7917	2.9583	2.4583
3.0000	2.9583	2.9583	2.9583	3.0000	2.9167	2.9583	2.9583

Table 12 Mann Whitney test between questionnaire LANA I and LANA II.

Test Statistics ^a								
	Enjoyable	Satisfaction with the information.	Remembering and understanding	Satisfaction with learning style	Design and colour	Learning from LANA is better than the teacher.	Using LANA again with other lessons.	Naturalness
Mann-Whitney U	288.000	288.000	288.000	276.000	276.000	240.000	288.000	144.000
Wilcoxon W	588.000	588.000	588.000	576.000	576.000	540.000	588.000	444.000
Z	.000	.000	.000	-.590	-1.000	-1.727	.000	-3.771
Asymp. Sig. (2-tailed)	1.000	1.000	1.000	.555	.317	.084	1.000	.000

a. Grouping Variable: Prototype

Table 12 illustrates the results of the Mann Whitney test conducted on the matching questions from the questionnaires administered in the questionnaire LANA I and LANA II evaluations. The results show that there are one significant differences in question 8 (the naturalness) p-value = 0.000163. The effect size (r) = 0.5 indicating a large effect size using Cohen (1988) criteria of 0.1 = small effect, 0.3 = medium effect, 0.5 = large effect. When comparing the mean values from questionnaire LANA I and LANA II it is found that four questions (Q1,Q2,Q3,Q7) from both questionnaires have the same mean values whereas the mean values of four questions (Q4,Q5,Q6,Q8) in questionnaire two are higher than questionnaire one which indicated an improvement in these questions rating.

Experiment 2: Adapting the VAK model to the autistic students learning Style

Participants

Thirty autistic students were recruited from five mainstream primary schools, that providing care for children with ASD in Saudi Arabia and Bahrain. Each child had previously received a community clinical diagnosis of high functioning autism spectrum disorder or Asperger's, who they have no problem with language and intellectual impairments. All participants with the age group (10-12) years old and their first language is Arabic and none of them had a previous experience using LANA II. An information sheet about the project and its aims and a consent form were sent to participant's parents to get their permission before conducting the experiment. The participants were divided into two equal sized groups. The first group was the Control Group (15 participants) used the LANAI without adapting to the learning style VAK model. The second group was the Experimental Group (15 participants) used LANA II with adapting to the learning style VAK model.

Experiment Methodology

This experiment was conducted to test the hypothesis H1_B based on the matrices outlined in Table 13. The data for this experiment is gathered from the LANA II log file and the user feedback questionnaire. The subjective and objective metrics were used to answer questions related to Hypothesis H1_B (LANA II can be adapted to the autistic student learning style): Question1: Can LANA II with the VAK have better results in the autistic student's perceptions? Question2: Can LANA II with the VAK learning Style improve the learning gain of the autistic students?

The tutorial that was given by LANA II is based on Science tutorials for autistic children age group (10-12) years old. All experiments were undertaken in a classroom environment, and each student individually worked on a computer. The computer used in this experiment is touchscreen computer in order to let the students use both of the keyboard and the screen to interact with LANA II. The teacher was present during the experiment and the students were given instructions about using LANA II.

During the tutorial session, all the questions from the system, the student answers, and the conversation variables/metrics were captured and saved in the log file, for further analysis to test the success of the tutorial session.

Table 13 Experiment 2 metrics

Metrics to be Evaluated	Method of Evaluation	Metrics type
Q1: LANA Tutor is enjoyable to use	Questionnaire	Subjective
Q2: LANA has a right amount of information in one Tutorial.	Questionnaire	Subjective
Q3: I remember what LANA has taught me in the tutorial.	Questionnaire	Subjective
Q4: I can learn better with using my learning style.	Questionnaire	Subjective
Q5: I like the design and colour of the interface.	Questionnaire	Subjective
Q6: I think I can learn from LANA better than my teacher.	Questionnaire	Subjective
Q7: I think I would like to use LANA again with other lessons.	Questionnaire	Subjective
Pre-test and Post-test	Log file	Objective

Experiment 2 Results

This section presents the collected results. The data was extracted from the tutoring log file (pre-test and post-test scores), and the questionnaire (see Table 13). These two data (pre-test and post-test scores, and Questionnaire) used to test the hypothesis H1_B (LANA II can be adapted to the autistic student learning style).

Pre-Test and Post-Test – Results

The learning gain was measured using a pre-test and post-test approach (Kelly and Tangney, 2006; Graesser et al., 2003; Lee et al., 2004). The same multiple-choice test questions were completed before and after the tutoring conversation. The test scores were compared to establish whether there is any improvement as follows:

$$\text{Learning gain} = \text{post_test} - \text{pre_test} \text{ (Eq 2)}$$

Table 14 shows the results of Wilcoxon Signed Ranks test carried out to determine if there was any difference between the pre-test and post-test scores with adapting to VAK leaning style model.

Table 14 Wilcoxon Signed Ranks test of the pre-test and post-test scores with adapting to the VAK

	N	Mean	Std. Deviation	Descriptive Statistics				
				Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
Pre-test	15	4.200	1.56753	1.00	7.00	3.0000	4.0000	5.0000
Post-test	15	9.800	.41404	9.00	10.00	10.0000	10.0000	10.0000
Test Statistics^a								
					Post-test - Pre-test			
Z					-3.431 ^b			
Asymp. Sig. (2-tailed)					.001			

When comparing the mean ranks in Table 14 of the two scores, the post-test scores was ranked higher, which means that the students performed better in the post-test. The p-values less than 0.05 ($p = .001$). The effect size (r) = 0.6 indicating a large effect size using Cohen (1988) criteria of 0.1 = small effect, 0.3 = medium effect, 0.5 = large effect. It can be concluded that there is a statistically significant difference between the pre-test and post-test scores with adapting to VAK learning style model. Table 15 illustrates the results of the Mann-Whitney U test conducted in order to measure whether there was a statistically significant difference in the learning gain between the two groups (Control Group: without adapting to VAK, Experimental Group: with adapting to VAK). From this result, it can be concluded that the learning gain between the Control Group and the Experimental Group are statistically significant p-value less than 0.05 ($P = .000009$). When comparing the mean ranks it can be seen that the Experimental Group has a higher rank than the Control Group, meaning that the learning gain for Experimental Group was significantly improved compared to the learning gain of the Control Group. The effect size (r) = 0.8 indicating a very large effect size using Cohen (1988) criteria of 0.1 = small effect, 0.3 = medium effect, 0.5 = large effect.

Table 15 Mann-Whitney U test for the learning gain between Control Group and Experimental Group.

	Group	N	Mean Rank	Sum of Ranks
Learning gain	Control	15	8.50	127.50
	Experimental	15	22.50	337.50
	Total	30		
				Learning gain
Mann-Whitney U				7.500
Wilcoxon W				127.500
Z				-4.450
Asymp. Sig. (2-tailed)				.000009

Relative learning gain was also measured in this experiment. Relative learning gain is a measure that calculate the average improvement in test scores as a percentage of the possible improvement (Latham et al., 2014). This measure additionally takes into account the opportunity for improvement. Average test score improvements were calculated and compared using the following formula:

$$\text{Relative learning gain} = (\text{PostTest} - \text{PreTest}) / (\text{TotalPossibleScore} - \text{PreTest}) \quad (\text{Eq3})$$

For example, if student get 8/10 in pre-test and only improves by 1, this is different to another getting 3/10 in pre-test and only improving by 1 – improvement is 50% in first case but only 14.2% in the second case.

Table 16 Mann-Whitney U test for the Relative learning gain between Control Group and Experimental Group.

Ranks				
	Group	N	Mean Rank	Sum of Ranks
Relative Learning gain	Control	15	8.00	120.00
	Experimental	15	23.00	345.00
	Total	30		
Test Statistics ^a				
	Learning gain			
Mann-Whitney U				.000
Wilcoxon W				120.000
Z				-4.838
Asymp. Sig. (2-tailed)				.000001
a. Grouping Variable: VAK				

Table 16 shows that there is a difference in the mean value between Control Group and Experimental Group. The average improvement in test scores in the Experimental Group were improved more than the Control Group. The ranks in each case are: Control Group (M= 8.00) and Experimental Group (M= 23.00). The learning gain between the Control Group and the Experimental Group are statistically significant different, p-value less than 0.05 (P= .000001). The effect size (r) is calculated using the (Eq1) and found that $r = 0.8$ indicating a large effect size using Cohen (1988) criteria of 0.1 = small effect, 0.3 = medium effect, 0.5 = large effect. It can be concluded that the average improvement in test scores in the Experimental Group are statistically significant comparing to the Control Group.

Adapting VAK model to autistics learning style (Questionnaire) – Results

Table 17 and Table 18 demonstrate the results of the third question in the questionnaire to test the participant's perception of remembering what LANA II has taught them between Control Group (without adapting to the VAK) and Experimental Group (with adapting to VAK).

Table 17 Crosstab *between Control Group and Experimental Group - Question3.

			Group		Total
			Experimental	Control	
I think I remember what LANA has taught me in the tutorial	Happy	Count	14	8	22
		% within VAK	93.3%	53.3%	73.3%
	Normal	Count	1	7	8
		% within VAK	6.7%	46.7%	26.7%
Total		Count	15	15	30
		% within VAK	100.0%	100.0%	100.0%

Table 18 Chi-Square Test ** between Control Group and Experimental Group - Question3.

Chi-Square Tests between Control Group and Experimental Group					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	6.136 ^a	1	.013		
Continuity Correction ^b	4.261	1	.039		
Likelihood Ratio	6.719	1	.010		
Fisher's Exact Test				.035	.018
N of Valid Cases	30				

Table 17 illustrates that the majority of students in the Experimental Group were remembering the tutorial (93.3%) and (6.7%) of the students have a neutral feeling; whereas students in the Control Group were remembering the tutorial with a percentage of (53.3%) and (46.7%) stated they have a neutral feeling when rating this question. A Chi-Square Test in Table 18 shows that there is statistically significant relationship between (adapting, not adapting) to the learning style and answering this question (P- value less than .05) $P=.013$. Table 19 shows the result of the fourth question in the questionnaire to test the participant's perception of using their learning style in the tutorial.

Table 19 Crosstab Question4: I think I can learn better with using my learning style (Visual – Auditory - Kinaesthetic).

		VAK		Total
		With		
I think I can learn better with using my learning style (Visual – Auditory - Kinaesthetic)	Happy	Count	14	14
		% within VAK	93.3%	93.3%
	Normal	Count	1	1
		% within VAK	6.7%	6.7%
	Total	Count	15	15
		% within VAK	100.0%	100.0%

The results from Table 19 highlighted that the majority of the students are happy with adapting to their learning style with percentage (93.3%), whereas (6.7%) of the student have a neutral feeling when rating this question. Table 20 and Table 21 illustrate the result of the sixth question in the questionnaire to test the participant's perception of learning from LANA II more than leaning from the teacher.

Table 20 Crosstab * between Control Group and Experimental Group - Question 6

		Group		Total
		Experimental	Control	
I think I can learn from LANA better than my teacher	Happy	Count	14	23
		% within VAK	93.3%	76.7%
	Normal	Count	1	7
		% within VAK	6.7%	40.0%
Total	Count	15	15	30
	% within VAK	100.0%	100.0%	100.0%

Table 21 Chi-Square Test * between Control Group and Experimental Group – Question 6.

Chi-Square Tests				
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	4.658 ^a	1	.031	
Continuity Correction^b	2.981	1	.084	
Likelihood Ratio	5.058	1	.025	
Fisher's Exact Test				.080
N of Valid Cases	30			

a. 2 cells (50.0%) have expected count less than 5. The minimum expected count is 3.50.

b. Computed only for a 2x2 table

The results show that the majority of students in the Experimental Group were preferring to learn from LANA II than learning from the teacher with percentage of (93.3%), and (6.7%) of the students have a neutral feeling. The students in the Control Group preferred to learn from LANA II than learning from the teacher with percentage of (53.3%), and (46.7%) stated they have a neutral feeling when rating this question. A Chi-Square Test in table 21 shows that there is statistically significant relationship between (the adapting, not adapting) to the learning style and answering this question (P- value less than .05) $P=.031$. The frequency analysis of the questionnaire results illustrated in Table 22 shows the participants satisfaction with the usability of LANA II. The results indicated that the majority (76.7%) of the participants felt that learning using LANA II is better than learning from the teacher in the class.

Table 22 Frequency analysis LANA II questionnaire

QUESTIONNAIRE RESULTS	Happy	Normal	Sad
I think LANA Tutor is enjoyable to use	100%	0	0
I think LANA had a right amount of information in one Tutorial.	91.6%	8.4%	0
I remember what LANA has taught me in the tutorial.	73.3%	26.7%	0
I can learn better with using my learning style.	93.3%	6.7%	0
I like the design and colour of the interface.	91.6%	8.4%	0
I think I can learn from LANA better than my teacher.	76.7%	23.3%	4.2%
I think I would like to use LANA again with other lessons.	79.16%	20.84%	0

Comparative descriptive analysis of data from autistic participants and neurotypical participants

This section presents further descriptive analysis of the collected data through the log file. The analysis of results presented in this part are intended to highlight differences of the results between autistic participants and neurotypical participants. The Man-Whitney U test was used in (Table 23 and Table 24) to highlight any statistically significant differences in the data.

Table 23 Mean rank values from Mann-Whitney U test for the objective data between autistic participants and neurotypical participants

	Participants	N	Mean Rank	Sum of Ranks
Average conversation duration (mins)	NT	24	12.50	300.00
	Autistics	30	39.50	1185.00
Number of unrecognised utterance	NT	24	31.63	759.00
	Autistic	30	24.20	726.00
Number of correct response	NT	24	22.81	547.50
	Autistic	30	31.25	937.50
Number of incorrect response	NT	24	32.38	777.00
	Autistic	30	23.60	708.00
Number of utterances using Similarity Strength	NT	24	39.54	949.00
	Autistic	30	17.87	536.00

Table 24 P- Values from the Mann-Whitney U test for objective data between autistic participants and neurotypical participants.

	Average conversation duration (mins)	Test Statistics ^a			
		Number of unrecognized utterance	Number of correct response	Number of incorrect response	Number of utterances using Similarity Strength
Mann-Whitney U	.000	261.000	247.500	243.000	71.000
Wilcoxon W	300.000	726.000	547.500	708.000	536.000
Z	-6.866	-2.002	-2.560	-2.625	-5.327
Asymp. Sig. (2-tailed)	.000	.045	.010	.009	.000
a. Grouping Variable: Prototype					

From Table 23 and Table 24, the results show that there was a statistically significant difference in the time taken to complete the tutorial between the two groups participants (reported by the SPSS as $p = .000$). When the mean ranks were compared, the results show that autistic participants rank higher than the neurotypical participants, indicating that the autistic participants spent time more than the neurotypical participants to complete the tutorial. Moreover, the results show that there was a statistically significant difference in the number of unrecognized utterance between the two groups participants ($p = .045$). When the mean ranks were compared the results show that neurotypical participants ranks higher than the autistic participants, indicating that the number of unrecognized utterances from neurotypical participants was more than the autistic participants.

When comparing the number of correct responses between the two groups participants, the results show that there was a statistically significant difference ($p = .010$). In comparing the mean rank the autistic participants rank higher, which indicates that the autistic participants having more correct responses than the neurotypical participants.

In addition, the results indicated that there were statistically significant differences between the numbers of the incorrect response between the autistic participants and the neurotypical participants ($p = 0.009$). When comparing the mean ranks it can be seen that the neurotypical participants ranks higher than autistic participants, meaning the neurotypical participants having more incorrect responses than the autistic participants.

Furthermore, there was a statistically significant difference between the two groups participants (reported by the SPSS as $p = .000$). When comparing the mean ranks it can be seen that the neurotypical participants ranks higher than autistic participants, meaning that the number of utterance used the similarity strength in neurotypical participants were more than the number of autistic participants, which means most of the autistic student's utterances were matched the scripted patterns.

DISCUSSION

When looking at the improvements made in the LANA II algorithm which were adopted to recognise common spelling variations, a morphologic pre-processing method was applied to address the common spelling variations and mistakes made by user because of the phonological similarity of some of Arabic alphabet characters. Furthermore, the similarity algorithm (i.e. Cosine similarity) was redeveloped to allow the common variations of certain words to be recognised and responded correctly. These improvements have reduced the number of unrecognised utterances and incorrect responses which are evident in the results. The evaluation results taken from the log file reveal that unrecognised utterances were reduced to 5.4% in LANA II compared to 38% in LANA I (Table 7). The results of the test revealed a statistically significant difference between the numbers of unrecognised utterances, with LANA II having a lower mean rank of unrecognised utterances. The results also showed that the incorrect responses were reduced to 3.6% (Table 7) in LANA II compared to 10.2% in LANA I. there are statistically significant differences (Table 8 and Table 9) between the numbers of the incorrect response between LANA I and LANA II ($p = 0.001$). In terms of the correct responses, the results reveal that the correct responses using the new STS algorithm in LANA II (96%) was significantly improved compared to the correct responses using STS algorithm of LANA I (89%). The findings of the questionnaire from the evaluation of LANA I and

LANA II show there was a significant difference in question 8 (the naturalness) $p\text{-value} = 0.000163$. This revealed that the users perceived LANA II to be better in relation to conversation naturalness with 91.6% of the participants expressing that the conversation level of naturalness was good which is a major improvement from LANA I where 45.8% who expressed the conversation level of naturalness was good. Hence based on these results there is enough evidence to suggest that H1_A (The enhancements made in LANA II architecture improve the overall effectiveness of LANA II engine) can be accepted.

When looking at the experiment with the autistic students, the evaluation results taken from the log file reveal that the pre-test and post-test scores between the Control Group (without adapting to the VAK) and the Experimental Group (with adapting to the VAK) are statistically significant $p\text{-value}$ less than 0.05 ($P = .000009$). In addition, the Experimental Group has a higher mean rank than the Control Group, meaning that the learning gain for the Experimental Group was significantly improved compared to the learning gain of the Control Group. The results from the questionnaire reveal the participant's perception of remembering what LANA II has taught them between the Control Group and the Experimental Group. The majority of students in the Experimental Group were remembering the tutorial 93.3% and 6.7% of the students have a neutral feeling; whereas students in the Control Group were remembering the tutorial with a percentage of 53.3% and 46.7% stated they have a neutral feeling when rating this question. The analysis results showed that there was statistically significant relationship between (the adapting, not adapting) to the learning style and answering this question ($P\text{-value}$ less than .05) $P = .013$, indicating that adapting to the VAK learning style helped the students to remember the tutorial. Furthermore, the evaluation results taken from the questionnaire reveal the participant's perception of adapting to their learning style in the tutorial. The majority of the students are happy with adapting to their learning style with percentage 93.3%, whereas 6.7% of the student have a neutral feeling when rating this question. When looking at the results of the participant's preference of learning from LANA II more than leaning from the teacher, the evaluation results reveal that the majority of students in the Experimental Group preferred to learn from LANA II than learning from the teacher with percentage of 93.3%, and 6.7% of the students have a neutral feeling. The students in the Control Group preferred to learn from LANA II than learning from the teacher with percentage of 53.3%, and 46.7% stated they have a neutral feeling when rating this question. In addition, there was statistically significant relationship between (the adapting, not adapting) to the learning style and answering this question ($P\text{-value}$ less than .05) $P = .031$. It can be concluded from the above results and analysis that LANA II using VAK model succeed to adapt to the autistics learning style. Therefore, the hypothesis H_B: LANA II using VAK model can be adapted to the Autistics learning style, can be accepted.

Additional findings of interest derived from the evaluation of LANA II were highlighted through the comparison of the data gathered from the autistic students and neurotypical students. The results show that there was a statistically significant difference in the number of unrecognized utterance between the two groups participants ($p = .045$). The neurotypical participants mean value ranks higher than the autistic participants, indicating that the number of unrecognized utterances from neurotypical participants was more than the autistic participants. In addition, the number of correct responses between the two participants was statistically significant ($p = .010$). The autistic participants had more correct responses than the neurotypical participants. Furthermore, the results indicated that there were statistically significant differences between the numbers of the incorrect response between the autistic participants and the neurotypical participants ($p = 0.009$), meaning the neurotypical participants having more incorrect responses than the autistic participants. Most prominent of the differences observed between the autistic students and the neurotypical students was in terms of the number of utterances required the new STS algorithm, the results reveal that the number of utterances used the similarity strength in neurotypical participants 42.7% are more than autistic participants 31.12%. The comparison of the data from the two participant's datasets illustrated that the utterances from the neurotypical participants contained significantly more instances where the utterance required the string similarity algorithm. The most of utterances from the autistic participants are matched the scripted patterns without need to apply the STS algorithm. A reason for this could be that the way that the autistic students speak and write. They tend to use a classical to modern Arabic language (formal language) on their speaking and writing, whereas the neurotypical students usually speak and write modern to colloquial Arabic language (informal language). The language used in the scripted patterns is modern Arabic language (formal language), therefore the autistic student's utterances matched the scripted patterns, whereas the neurotypical students used some colloquial language in their writing, therefore their utterances didn't match the scripted patterns and required the string similarity algorithm.

CONCLUSION AND FUTURE WORK

This paper presented research into the development of an Arabic Conversational Intelligent Tutoring System for autistic children (LANA II). The research is part of investigation in to several key areas such as, Conversation Agent (CA), Intelligent tutoring system (ITS), Arabic Language Processing techniques (i.e. natural language processing, sentence similarities measures and pattern matching), learning style, TEACCH method for children with autism spectrum disorder. The main contribution in this paper is presenting a new Arabic CITS architecture called LANA II, which adapts the TEACCH method that is used to educate autistic children and adapts the VAK model to autistic students learning style. In addition, this paper describes the development of a new Arabic CA algorithm that reduces the scripting effort and unrecognised utterances by processing the spelling mistakes and morphological nature of the Arabic language, which were issues in LANA I. Moreover, a new general evaluation framework has been developed and tested from the objective and subjective perspective.

LANA II contains of two main components: the first component was the CA engine. CA algorithm was developed in order to reduce the effort required in scripting the knowledge base/domain by using pre-processing feature and new version of Cosine similarity approach. The second component of LANA II is the intelligent tutoring system which adapts the tutorial to the autistic children learning style. LANA II personalises the tutorial according to the children learning style: Visual, Auditory, Kinaesthetic (VAK). In addition, the tutorial interface is mapped to the TEACCH method, which is based on Structured Teaching technique for autistic students. The results of the log file and end user evaluation for (LANA I) revealed some weaknesses and some negative perceptions from the participants (S. Aljameel et al., 2018). The issues that were highlighted in the evaluation of LANA I were mainly due to grammatical features and the morphological nature of the Arabic language. Other issues that were identified were the need to further expand the knowledge base in order to increase the naturalness of the conversation. The participants expressed that they perceived the naturalness of their conversation with LANA I to be low. LANA II aims to overcome these weaknesses. From the evaluation, it is illustrated statistically significant improvements in terms of the objective and subjective metrics measured in the LANA II engine. In addition, LANA II using VAK learning style succeed to adapted to the autistics learning style and enhance their learning base on the objective and subjective metrics from the experiment. Inevitably as with any research and evaluation effort the evaluation of LANA II did highlight some areas which can be improved through further research and development such as the suggestions below:

1. Semantic similarity: LANA II similarity algorithm can be strengthened with the addition of semantic rather than lexical similarity. The paraphrased version of a scripted pattern will be recognised by using the semantic similarity. This will further reduce the scripting patterns and making the task of scripting an Arabic CA even less exhausting.
2. Knowledge base expansion: As the knowledge base of LANA II is based on the science subjects, specifically two lessons from the science book of grade 4, future work can entail the expansion of the knowledge base to cover all lessons of the science topics for grade 4. Moreover, the knowledge base could be expanded to cover other lessons for other grades. Furthermore, the general topics of the knowledge base could also be expanded to include more topics and patterns related to general topics.
3. Automatically predict autistic children learning styles: As detecting the learning style in LANA II is based on a questionnaire. However, if the questionnaire is not completed accurately, this leading to incorrect assessment of learning styles. Future work can be made to detect the autistic children learning style by analysing learner behaviour throughout tutoring.

References

- Aljameel, S., O'Shea, J., Crocket, K., Latham, A. and Kaleem, M. (2018) LANA: An Arabic Conversational Intelligent Tutoring System for Children with ASD. IEEE Transection on Learning Technologies.
- Aljameel, S. S., O'Shea, J. D., Crockett, K. A. and Latham, A. (2016) *Survey of string similarity approaches and the challenging faced by the Arabic language*. IEEE.

- Aljameel, S. S., O'Shea, J. D., Crockett, K. A., Latham, A. and Kaleem, M. (2017) *Development of An Arabic Conversational Intelligent Tutoring System for Education of Children with ASD*. IEEE.
- Alobaidi, O. G., Crockett, K. A., O'Shea, J. D. and Jarad, T. M. (2013) *Abdullah: An Intelligent Arabic Conversational Tutoring System for Modern Islamic Education*. Vol. 2:
- Barnhart, R. (2017) *A Phenomenological Study of Elementary Teachers of Autism Spectrum Disorder Students: Common Lived Experiences*. University of Louisiana at Lafayette.
- Boucenna, S., Narzisi, A., Tilmont, E., Muratori, F., Pioggia, G., Cohen, D. and Chetouani, M. (2014) 'Interactive technologies for autistic children: A review.' *Cognitive Computation*, 6(4) pp. 722-740.
- Dermeval, D., Paiva, R., Bittencourt, I. I., Vassileva, J. and Borges, D. (2017) 'Authoring Tools for Designing Intelligent Tutoring Systems: a Systematic Review of the Literature.' *International Journal of Artificial Intelligence in Education*, pp. 1-49.
- Dzikovska, M., Steinhäuser, N., Farrow, E., Moore, J. and Campbell, G. (2014) 'BEETLE II: Deep natural language understanding and automatic feedback generation for intelligent tutoring in basic electricity and electronics.' *International Journal of Artificial Intelligence in Education*, 24(3) pp. 284-332.
- D'Elia, L., Valeri, G., Sonnino, F., Fontana, I., Mammone, A. and Vicari, S. (2014) 'A longitudinal study of the TEACCH program in different settings: The potential benefits of low intensity intervention in preschool children with autism spectrum disorder.' *Journal of autism and developmental disorders*, 44(3) pp. 615-626.
- Fletcher-Watson, S. (2014) 'A targeted review of computer-assisted learning for people with autism spectrum disorder: Towards a consistent methodology.' *Review Journal of Autism and Developmental Disorders*, 1(2) pp. 87-100.
- Graesser, A. C. (2016) 'Conversations with AutoTutor help students learn.' *International Journal of Artificial Intelligence in Education*, 26(1) pp. 124-132.
- Grynszpan, O., Weiss, P. L., Perez-Diaz, F. and Gal, E. (2014) 'Innovative technology-based interventions for autism spectrum disorders: a meta-analysis.' *Autism*, 18(4) pp. 346-361.
- Kondrak, G. (2005) *N-gram similarity and distance*. Springer.
- Latham, A., Crockett, K. and McLean, D. (2014) 'An adaptation algorithm for an intelligent natural language tutoring system.' *Computers & Education*, 71 pp. 97-110.
- Lindsay, S., Proulx, M., Scott, H. and Thomson, N. (2014) 'Exploring teachers' strategies for including children with autism spectrum disorder in mainstream classrooms.' *International Journal of Inclusive Education*, 18(2) pp. 101-122.
- Massaro, D. W. and Bosseler, A. (2006) 'Read my lips: The importance of the face in a computer-animated tutor for vocabulary learning by children with autism.' *Autism*, 10(5) pp. 495-510.
- Meng, Q. (2017) Interventions promoting academic engagement for children with ASD in general education classrooms: A systematic literature review.
- Mesibov, G. B. and Shea, V. (2010) The TEACCH program in the era of evidence-based practice.' *Journal of autism and developmental disorders*, 40(5) pp. 570-579.
- Mesibov, G. B., Howley, M. and NetLibrary, I. (2003) *Accessing the curriculum for pupils with autistic spectrum disorders: using the TEACCH programme to help inclusion*. London: David Fulton Publishers.
- Parsons, S. and Cobb, S. (2011) 'State-of-the-art of virtual reality technologies for children on the autism spectrum.' *European Journal of Special Needs Education*, 26(3) pp. 355-366.

- Perikos, I., Grivokostopoulou, F. and Hatzilygeroudis, I. (2017) 'Assistance and feedback mechanism in an intelligent tutoring system for teaching conversion of natural language into logic.' *International Journal of Artificial Intelligence in Education*, 27(3) pp. 475-514.
- Qian, G., Sural, S., Gu, Y. and Pramanik, S. (2004) *Similarity between Euclidean and cosine angle distance for nearest neighbor queries*. ACM.
- Reynolds-Keefer, L. and Johnson, R. (2011) 'Is a picture is worth a thousand words? Creating effective questionnaires with pictures.' *Practical Assessment, Research & Evaluation*, 16(8) p. 2.
- Roberts, J. and Simpson, K. (2016) 'A review of research into stakeholder perspectives on inclusion of students with autism in mainstream schools.' *International Journal of Inclusive Education*, 20(10) pp. 1084-1096.
- Waddington, E. M. and Reed, P. (2017) 'Comparison of the effects of mainstream and special school on National Curriculum outcomes in children with autism spectrum disorder: an archive-based analysis.' *Journal of Research in Special Educational Needs*, 17(2) pp. 132-142.
- Yahaya, W. A. J. W. and Salam, S. N. A. (2016) 'Smiley Faces: Scales Measurement for Children Assessment.'